THE PHYSICAL SOCIETY

Reports on
PROGRESS IN PHYSICS

VOLUME IV

GENERAL EDITOR
ALLAN FERGUSON

1937

Reprinted 1967 for
Wm. DAWSON & SONS LTD., LONDON
with the permission of
THE PHYSICAL SOCIETY

Originally printed in Great Britain by the Cambridge University Press
Reprinted in Switzerland by Lang Druck Ltd., Berne

CONTENTS

PAGE

The measurement of time. By H. Spencer Jones 1

The adsorption of gases by solids. By Eric K. Rideal 27

Surface tension. By R. C. Brown 40

Sound. By E. G. Richardson 53

Supersonics in relation to molecular constitution. By E. G. Richardson . 67

Heat. By R. W. Powell 76

Thermodynamics. By J. H. Awbery 103

Refrigeration. By Ezer Griffiths 124

The beginnings of the new quantum theory. By H. T. Flint 134

Atomic physics. By C. H. Collie 173

Slow neutrons. By P. B. Moon 198

The charge of the electron. By Professor H. R. Robinson 212

Experimental electricity and magnetism. *Editor*, L. Hartshorn . . 218

§ 1. The breakdown of insulating materials. Electric strength. By L. Hartshorn
§ 2. Measurement of current at radio frequencies. By T. Iorwerth Jones
§ 3. Electrical oscillations and biological materials. By W. H. Ward
§ 4. Frictional electricity. By W. H. Ward

Electrolytes and electrolysis. By R. W. Gurney 251

Physical-optical instruments and materials. By F. Simeon 283

Spectroscopy. *Editor*, W. Jevons 295

§ 1. Spectroscopic methods of chemical analysis. By W. R. S. Garton
§ 2. Astronomical spectroscopy. By A. Hunter
§ 3. Vacuum spectroscopy. By R. W. B. Pearse
§ 4. Complex atomic spectra. By E. W. Foster
§ 5. Infra-red spectra of organic molecules. By Lotte Kellner

X-ray analysis and application of Fourier series methods to molecular structures. By J. Monteath Robertson 332

Diamagnetic and paramagnetic anisotropy of crystals. By Kathleen Lonsdale 368

PREFACE

THE appearance of the fourth annual volume of the *Reports on Progress in Physics*, issued under the auspices of the Physical Society, may fitly be made the occasion for a preface which shall say something of the aims and objects of the reports.

The policy of the Society is not fixed and immutable but is rather guided by the expression of the public needs and especially by the expression of the needs of the Fellows. The original suggestion was that the reports should give an annual conspectus of the progress made in the main divisions of physical science during the twelve months preceding the appearance of the current volume of the reports, and this aim has always been kept in mind, though the space devoted to each of those divisions varies with the interest and variety of the advances made in it. With each volume of these reviews were to be issued one or two articles dealing with prior developments, and not confined to recent advances, in some special subject. In the later volumes these articles have tended to increase in number and importance, a tendency which is certainly to be welcomed. The scope of the various reports is difficult to define and they naturally vary in standard and in outlook. They are designed to subserve two ends, not mutually conflicting. On one hand, a report may assist a specialist by telling him an ordered story of the most recent advances in his particular branch of physics. On the other hand, a trained physicist who is not a specialist in a certain branch of physics may desire to know something of the happenings therein, and it is sometimes possible to meet his needs as well as those of the specialist if a report does not plunge too suddenly *in medias res* but is prefaced by a section which deals with general principles and earlier developments, so leading up to the more difficult technical details. In some instances the reports are written wholly for the non-specialist physicist.

There is, on occasion, a certain amount of overlap between the contents of two articles and this, when it expresses a difference of outlook or of treatment, is not, in the Editor's view, to be deprecated.

Nor is it a matter for grave concern if an article on recent advances in nuclear physics appears simultaneously with an article having the same title and published under the auspices of, say, the Zetetic Society. The angle of view of the two contributors is likely to be different enough to give the two articles independent interest, and in any case the present reports are published primarily in the interests of the Fellows of the Society, and it is by a happy chance that they interest a public outside the membership of the Society. There is little need to waste time or sympathy on the plight of the busy and harassed researcher who intones the blessed word "co-ordination" with the reverence assigned to "Mesopotamia", and who knows not which of the identically-titled articles to select for the honour of his

attention or which to read first. "Sir," said Dr Johnson, "it no more matters which subject you teach your boy first, than which leg you put in your breeches first. Sir, you may stand disputing which leg were better to go in your breeches first, but meantime your breech is bare. Sir, while you are arguing which subject were best taught first to your boy, another boy will have learnt both."

The moral is obvious.

ALLAN FERGUSON

THE MEASUREMENT OF TIME

By H. SPENCER JONES, M.A., Sc.D., F.R.S., Astronomer Royal

§ 1. GENERAL CONSIDERATIONS AND DEFINITIONS

THE measurement of time depends fundamentally upon the rotation of the earth on its axis. This rotation provides a natural clock, to which all measurements of time must be referred. The sun, the moon, a planet or the stars may serve as the hands of this clock and according to the choice that is made, different times will result. Whichever body is selected, we may define the interval between two successive transits of that body over a given meridian as a day, and adopt the day, defined in this way, as the unit for the measurement of time.

The most natural unit of time is obtained by using the sun as a reference body. The time so determined is called *apparent solar time*. It is the time that is given by any form of sundial. The *apparent solar day* begins and ends at apparent midnight, the moment of lower meridian transit of the centre of the sun. The apparent solar time at any instant is therefore the westward hour angle of the true sun + 12 hr.

Our daily life must be regulated by solar time, but apparent solar time suffers from the disadvantage that the days are not of constant length. The inequality arises from two causes: (1) the orbit of the earth around the sun is not circular, so that the angular rate of apparent motion of the sun as seen from the earth is not uniform; (2) the sun moves in the ecliptic and not in the equator.

A fictitious mean sun is therefore imagined, which moves in the celestial equator at a uniform rate and completes its passage round the equator in the time that the true sun takes to pass round the ecliptic. Then the time given by this fictitious sun will be such that every day is of the same length and equal to the average length of the apparent solar day. The interval between successive transits of this mean sun constitutes the *mean solar day*. The mean sun is started, relatively to the true sun, according to a certain convention—which need not be explained here—so that *mean solar time* never differs widely from apparent solar time.

The mean sun crosses different meridians at different moments, so that for any place there is a *local mean time*, the local mean noon being defined as the passage of the mean sun across the meridian passing through that place. It is convenient to adopt one meridian as a standard, to which all other mean times are referred. By international agreement, based on the recommendation of the Prime Meridian Conference, which met in Washington in 1884, the meridian of Greenwich has been adopted as the prime meridian. *Greenwich mean time* is the standard to which all other mean times are referred. It is now generally called *universal time* (U.T.), or in German *Weltzeit* (W.Z.). For any place in longitude λ, measured westward from the Greenwich meridian, we have therefore

$$\text{Local mean time} = \text{universal time} - \lambda.$$

To avoid the inconvenience of using numerous local mean times, a system of standard or zone times has been adopted by most of the principal countries of the world. Usually the standard times differ from Greenwich mean time by an integral number of hours.

The difference between mean time and apparent time is called the *equation of time*. This quantity is also the difference between the hour angles of the mean and true suns. The equation of time vanishes four times in a year, on or about 15 April, 15 June, 31 August and 24 December. Its maximum positive value is nearly $16\frac{1}{2}$ min., about 3 November, and its maximum negative value is nearly $14\frac{1}{2}$ min., about 12 February.

If, instead of the sun, the stars are used as the hands of the clock we obtain *sidereal time*. The *sidereal day* can be defined as the interval between two successive transits across the meridian of any place of a star (assumed to have no proper motion). This interval is $23^h\ 56^m\ 4^s{\cdot}100$ of mean solar time and is equal to the period of the axial rotation of the earth. In practice, however, the transits of the first point of Aries or vernal equinox (the point where the ecliptic passes from south to north through the equator) are used to define the sidereal day. The transit of the first point of Aries determines the beginning of the sidereal day, so that sidereal time may be defined as the hour angle of the vernal equinox.

Because of the annual retrograde motion of the equinox along the ecliptic, amounting to 50 sec. and called *precession*, the adopted sidereal day is 0·009 sec. shorter than the true period of rotation of the earth and corresponds to $23^h\ 56^m\ 4^s{\cdot}091$ of mean solar time. Conversely, the mean solar day is $24^h\ 3^m\ 56^s{\cdot}555$ of sidereal time.

The *right ascension* of any celestial object is defined as the interval, measured in sidereal time, between the transit of the vernal equinox and the transit of the body. The sidereal time of the meridian transit of a body is therefore equal to its right ascension. In particular, the sidereal time at the moment of the meridian passage of the mean sun, which is the moment of mean noon, is equal to the right ascension of the mean sun. A knowledge of the right ascension of the mean sun would therefore permit the conversion of mean into sidereal time or of sidereal into mean time.

We have mentioned that the right ascension of the mean sun is measured from the true equinox. But the precessional motion of the true equinox is not uniform, being affected by irregularities, due to solar and lunar perturbations, which are known as *nutation*. In consequence, the sidereal time, as we have defined it, does not increase uniformly and the sidereal day varies in length. These variations have become of importance during the past decade as a result of the improved time-keeping of precision clocks. By analogy with apparent solar time, we may therefore speak of the sidereal time defined above as *apparent sidereal time*.

If we imagine a point to move uniformly along the equator, with a motion equal to the mean motion of the true equinox, and so that its extreme distances from the true equinox on both sides are equal, we may term this point the *mean equinox*. The right ascension of the mean equinox, measured from the true equinox, is the

nutation in right ascension, and has the extreme values $\pm 1{\cdot}2$ sec. We can define *mean sidereal time* by reference to this mean equinox, so that

Mean sidereal time = Apparent sidereal time − nutation.

Mean sidereal time, as thus defined, increases uniformly and all sidereal days have the same length. Observations of transits of stars with the transit circle determine *apparent sidereal time*.

We have therefore the following analogy:

	Solar	Sidereal
Observations give	Apparent solar time	Apparent sidereal time
Perfect clocks keep	Mean solar time	Mean sidereal time
The difference is	Equation of time	Nutation in right ascension

The nutation in right ascension is tabulated for 0 hr. each day in the *Nautical Almanac*. It comprises terms of long period, depending upon the longitude of the moon's node, the longitude of the sun and the longitude of the sun's perigee, which have a principal period of 18 years and a range of $\pm 1{\cdot}2$ sec., and terms of short period, which have a principal period of nearly 15 days and a range of $\pm 0{\cdot}020$ sec.

The values of the nutation in right ascension for 0 hr. on the first day of each month during 1938 are given in table 1.

Table 1

		sec.			sec.
1938	1 Jan.	+0·988	1938	1 Aug.	+0·924
	1 Feb.	+1·037		1 Sept.	+0·877
	1 Mar.	+1·000		1 Oct.	+0·793
	1 Apr.	+0·890		1 Nov.	+0·723
	1 May	+0·822		1 Dec.	+0·718
	1 June	+0·850	1939	1 Jan.	+0·767
	1 July	+0·909			

The above figures emphasize the necessity of using mean sidereal time instead of apparent sidereal time when the performance of a precision timekeeper is to be checked by means of observation.

§ 2. METHODS OF DETERMINING TIME

The transit instrument. Time is usually determined by observations of the instants of the transits of stars across the meridian, using a transit instrument. When a star is on the meridian, the apparent sidereal time is equal to its right ascension. If, then, the clock time of the transit is determined, the error of the clock can be found provided that the right ascension of the star is known.

The transit instrument consists essentially of a refracting telescope, which is supported at the ends of an axis, perpendicular to the tube, by two trunnions moving in Y bearings. The axis is horizontal and points east and west so that the telescope, as it swings on the axis, moves in the meridian. A graduated circle is fixed to the axis, enabling the instrument to be set on a star of any required

declination. In the focal plane of the object-glass is placed a framework carrying a number of vertical wires (spider-webs) and a single horizontal wire across the centre of the field.

Observations were originally made by the eye-and-ear method, the observer watching the star and listening to the beats of the clock, interpolating the times of transit across each wire to the nearest tenth of a second. This method has been entirely superseded by electrical registration. Until about 25 years ago, it was customary to use a hand tapper, by means of which the observer momentarily closed an electric circuit as the star appeared to him to cross each wire. The tap was recorded on a chronograph, together with signals sent each second by the clock.

Both methods of observation, the eye-and-ear and the tapping method, are affected by two types of error, which enter directly into the determination of time, viz. *personal equation* and *magnitude equation*.

Systematic differences between two observers using the eye-and-ear method were first noticed by Maskelyne in 1794, as shown by the following quotation[1]:

"I think it necessary to mention that my assistant, Mr David Kinnebrook, who had observed the transits of the stars and planets very well, in agreement with me, all the year 1794, and for the great part of the present year, began, from the beginning of August last, to set them down half a second of time later than he should do, according to my observations, and in January of the succeeding year, 1796, he increased his error to $\frac{8}{10}$ths of a second. As he had unfortunately continued a considerable time in this error before I noticed it, and did not seem to me likely ever to get over it, and return to a right method of observing, therefore, though with reluctance, as he was a diligent and useful assistant to me in other respects, I parted with him. The error was discovered from the daily rate of the clock deduced from a star observed on one of two days by him, and on the other day by myself, coming out different from what it did from another star observed both days by the same person, either him or myself....The method of observing, introduced by Bradley, of noting the proportional distance of the star from the wire at the two beats immediately preceding and following the transit across the wire, has been carefully used, and thence the transit set down to the tenth of a second. I cannot persuade myself that my late assistant continued in the use of this excellent method of observing, but rather suppose he fell into some irregular and confused method of his own; as I do not see how he could have otherwise committed such gross errors."

When the tapping method is used, it is easy to see how personal differences can arise. One observer may wait until he sees the star bisected by the wire and then make his tap; another observer may, as it were, shoot the flying bird. The first observer will be systematically later than the second. The systematic differences between various observers may be large: thus, fifteen observers at Greenwich showed an extreme range of 0·7 sec.[2]. The differences between experienced observers remained sensibly constant, provided that the observer did not change, consciously or unconsciously, his method of observing. Considering, for instance, the three

years, 1912–14, there were seven observers at Greenwich who participated in the transit circle observations for each of these three years. The mean yearly differences from the standard observer, W., of the other six observers were as shown in table 2.

Table 2. Unit, 0·01 sec.

Observers	1912	1913	1914
W.–B.	+11	+14	+17
W.–W.B.	+23	+27	+24
W.–H.F.	+19	+13	+ 7
W.–R.C.	−13	−10	−16
W.–H.A.	+34	+32	+32
W.–W.D.	−13	− 7	− 7

The personal equations, though approximately constant in the mean, are liable to vary somewhat from night to night, possibly depending upon the conditions of seeing and the size of the star images, and are liable also to change during a long spell of observing as the observer becomes fatigued.

The *magnitude equation* is a systematic effect depending on the magnitude, or brightness, of the star observed, faint stars being observed systematically late as compared with bright stars. This systematic effect had existed unsuspected for many years before it was discovered by Gill, from a comparison between a triangulation with the heliometer of certain stars, which were used as comparison stars for observations of Mars made at Ascension in 1877, with meridian observations made at several observatories[3]. Gill's explanation was "that in chronographic registration the tendency will be to press the key too late for faint stars, because their small near disks make them appear farther from the web than a bright star with its larger disk and rings at the same angular distance from the web". Gill thought, at first, that observations by the eye-and-ear method were relatively free from this source of error, but it was subsequently found that they are also affected in a similar manner.

Table 3 shows the mean magnitude equations for a series of Greenwich catalogues, expressed as corrections to be applied to the catalogue right ascensions for the stellar magnitudes given in the first column[4]:

Table 3. Unit, 0·001 sec.

Mag. \ Cat.	1815	1830	1850	1860	1864	1872	1880	1890	1900	1910
1	+18	+22	+ 8	+15	+20	+20	+19	+13	+12	...
4	− 4	− 4	− 2	− 3	− 4	− 4	− 4	− 3	− 2	− 2
5	−11	−13	− 5	− 9	−12	−12	−11	− 8	− 7	− 6
6	−18	−22	− 8	−15	−20	−20	−19	−13	−12	−11
7	−26	−31	−12	−21	−28	−28	−26	−18	−16	−15
8	−33	−40	−15	−27	−36	−36	−34	−23	−21	−19

The magnitude equation is different for different observers and both personal equation and magnitude equation depend, for a given observer, upon the rate of

motion of the star or, in other words, they vary with declination. The personal equation depends, further, upon the direction of motion (right or left) of the star across the field; by using a reversing prism in front of the eyepiece, and reversing the apparent direction of motion of the star in the middle of the transit, this effect can be eliminated.

The impersonal micrometer. It was suggested by J. Repsold[5], in 1889, that the personal and magnitude equations could be eliminated by using a form of micrometer in which the web is made to follow the image of the star during its transit by motions communicated by the observer's hands alone. The axis, which imparts motion to the slide on which the travelling web is mounted, is provided with a disk at each end, so that the observer can use the thumb and finger of both hands in rotating it. After a little practice, there is no difficulty in keeping the web constantly bisecting the star. The drum of the micrometer screw is provided with electric contacts, so that during the transit the instants when the star is at known intervals from the line of collimation are automatically registered, along with the clock signals. A detailed description of two forms of eye-end for a transit circle, fitted with means for observing in this manner, were described later by Repsold[6], who gave the name of impersonal micrometer to this form of micrometer eye-end. A modification of the method was proposed by Struve[7], who suggested the use of clockwork to drive a cone at uniform speed, the cone driving by friction a disk, which turns an axis that transmits the motion to the micrometer screw. The speed of motion of the travelling wire could be varied, to suit the declination of the star under observation, by sliding the disk along the cone.

The impersonal micrometer of the Cape reversible transit instrument was fully described by Gill[8].

The impersonal micrometer, with or without the auxiliary drive, is almost universally used for the accurate determination of time and of right ascensions. The auxiliary drive reduces the strain on the observer, as it leaves him only the task of perfecting the guiding, and it undoubtedly reduces the accidental errors of observation. But with small instruments, which can be reversed during each transit, it is doubtful whether this additional complication is justified.

The personal equation of the observer is very much reduced by the use of the impersonal micrometer, but it is not entirely eliminated. Small systematic differences, which rarely exceed $\pm 0{\cdot}02$ sec., persist. These small differences remain extremely constant for considerable periods of time. Table 4, for instance, gives

Table 4. Unit, 0·001 sec.

Observer	1918–19	1920–21	1922–23
R.C.	+15	+17	+17
A.W.	−12	−16	−24
H.P.	− 5	− 3	0
M.	+16	+16	+15
D.	...	− 7	−18
L.D.	...	−12	− 6

the relative personal equations of several observers at the Cape Observatory for three consecutive periods of two years[9].

Small though these personal equations are, when compared with those that occur with the tapping method of observation, they are not negligible for the accurate determination of time. The internal probable error of a good time determination with a small transit instrument is of the order of $\pm 0{\cdot}01$ sec. so that, where great accuracy is desired, it is necessary to determine the absolute personal equation of the observer.

This may be done with a *personal equation machine*. In such a machine an artificial star is moved at a uniform speed by means of an accurate screw. The transit of this artificial star is observed in the transit instrument in the same way that a real star transit is observed, the instrumental contacts being recorded on a chronograph together with the signals from the standard clock. The personal equation machine carries a series of contacts and the times of the passages of the artificial star across these are also recorded on the chronograph. The true time of the passage of the artificial star across the middle of the contacts of the apparatus is thus compared with the observed time of its passage across the middle of the contacts of the telescope. When the artificial star reaches the end of its traverse, its direction is automatically reversed and the transit in the reverse direction is observed. By means of a suitable variable gearing, the artificial star may be made to transit at the rate appropriate to an equatorial star or at slower speeds corresponding to stars of higher declination. Observations are made both without and with a reversing prism over the eyepiece.

The personal equation appears to be due to two main causes: (*a*) bisection error; (*b*) following error.

Bisection error. An error of this nature occurs when the observer always sets the moving wire to the right or to the left of the centre of the star image, irrespective of the direction of motion of the star in the field of the telescope. In most transit instruments the observer faces in opposite directions when observing north and south stars respectively; the bisection error consequently changes sign at the zenith. In the broken transit, however, the observer always looks along the axis of the pivots and there is no discontinuity at the zenith. It will have the same sign, in the former case, for clock stars that transit to the south of the zenith as for circumpolar stars at lower transit; it will also have the same sign (but opposite to the preceding) for clock stars that transit to the north of the zenith as for circumpolar stars at upper transit. If it can be assumed that the angular error of bisection is independent of the speed of motion of the star, the effect on the time of transit will be proportional to sec δ. There is some evidence, however, that the angular error depends upon the speed of motion.

Following error. An error of this nature occurs if the observer, when observing a moving star image, systematically sets the wire either in front of or behind the centre of the image of the star. This type of error does not change sign at the zenith. It has the same sign for circumpolar stars at lower transit as for clock stars and also for circumpolar stars at upper transit. This error will also vary proportionally

to sec δ, if the angular error of setting is independent of the apparent speed of the star.

For the majority of observers, the bisection error is larger than the following error. The bisection error at the zenith can be determined by observing transits of stars within a few degrees of the zenith, one-half of each transit being observed with the observer facing north and the other half with the observer facing south.

Instrumental errors. Ideally, the axis of the transit instrument should lie in a horizontal plane and point due east and west. Small errors of level and azimuth are always present, however, and must be determined and allowed for. If the axis is inclined to the horizon by a small angle b, the east end being the lower, and also by the small angle k to the north of east and if there is an error of collimation, c, the resulting error in the observed time of transit of a star of declination δ is

$$t = b \cos(\phi - \delta) \sec \delta + k \sin(\phi - \delta) \sec \delta + c \sec \delta,$$

where ϕ denotes the latitude.

This expression is generally written in the form

$$t = m + n \tan \delta + c \sec \delta,$$

where $m = b \cos \phi + k \sin \phi$, $n = b \sin \phi - k \cos \phi$.

Small transit instruments are usually designed so that they can be reversed rapidly. It is then possible to reverse the instrument in the middle of each transit. As the sign of the collimation error is changed by reversing the instrument, the collimation error is eliminated by taking the mean of the two halves of the transit. We then have

$$t = m + n \tan \delta.$$

If the differences between the observed and tabular times of transit are plotted against the values of tan δ a series of points are obtained which should lie closely along a straight line. Errors of observation and of the tabular places of the stars will cause a certain scatter. The slope of the line determines n and its intercept on the axis tan δ = 0 determines m. This procedure is only legitimate if there is no change in level or azimuth during the observations. With small instruments there is usually a change in the level with change of temperature; the level error should be determined at short intervals with a spirit level and a curve drawn to represent the variation of level during the observations. The error of azimuth can then be derived by plotting the differences between the observed and tabular times of transit for each star against the azimuth factor, sin $(\phi - \delta)$ sec δ.

Errors of the pivots. Errors in the observed times of transit, which depend upon the position of the instrument (i.e. circle E or circle W) and upon the zenith distance, will be produced if the pivots are not truly cylindrical in shape. These errors may easily amount to a few hundredths of a second of time for clock stars and may be appreciably longer for close circumpolars. The determination of level error by means of a striding level will also be affected by errors in the shape of the pivots. With large instruments it is usual to determine the pivot errors accurately and to apply corrections for them. With small instruments it is preferable to correct the figure of the pivots, using a Krupp mikrotast gauge for testing, until the

departure of the pivots from the cylindrical shape is so small that the corrections become negligible.

The photographic zenith tube. With a small reversible transit instrument the internal probable error of a time determination is not much greater than 0·01 sec. The clock correction deduced from a night's observations may at times stand out as much as 0·04 or 0·05 sec. from the results of adjacent nights, though the internal consistency of the observations may be all that can be desired. With the improvement in time-keeping of astronomical clocks, better determinations of time are needed in order to control their irregular fluctuations. An interesting step in this direction has been taken at the Washington Observatory, by adapting the photographic zenith tube for the determination of time. In this method of determining time, the personal equation of an observer is avoided and effects due to errors in the adopted pivot errors as well as of level and azimuth errors are eliminated. It is the accurate determination of level that offers the greatest difficulty in observations with a transit instrument.

The photographic zenith tube in use at Washington was designed by F. E. Ross, who has given a full description of it[10]. The essential parts of the instrument comprise an objective, which can be rotated in its own plane round a vertical axis; a mercury surface placed below the objective at a distance equal to one-half its focal length and a photographic plate placed immediately beneath the objective and mounted so that it is rigidly connected with the cell of the objective. The plate carriage is moved by clockwork to follow the diurnal motion of the star, so that an untrailed image of the star is formed. For the determination of time[11], the moving plate registers on a chronograph the instants when it reaches certain points in its movement. Three exposures, each of fifteen seconds, are given, the last of which terminates about five seconds before the meridian passage of the star. The rotary carrying the objective and plate carriage is then turned through 180°. In the second half of the observation, the plate moves back over the same course and the time record registers the times when it is at the same points as were recorded on its forward movement. The plates are measured in a comparator and the time of meridian passage of each star is deduced.

The results of time determinations by this method and by the use of two transit instruments, Prin No. 1 and Prin No. 2, have been given by Littell[11]. The average and maximum residuals from the adopted clock curve and the internal probable errors were as shown in table 5.

Table 5. Unit, 1 sec.

Instrument	Average residual error	Maximum residual error	Internal probable error
Zenith tube	±0·0063	0·017	±0·0068
Transit Prin No. 1	±0·0116	0·037	±0·0199
Transit Prin No. 2	±0·0114	0·034	±0·0156

These figures show clearly the superior consistency of the observations with the photographic zenith tube. There were, however, systematic differences between

the zenith tube and the transit instruments of the order of 0·02 to 0·03 sec., for which no explanation has been found. The zenith tube method necessitates the use of stars that are much fainter than the standard clock stars and whose positions are not known with such accuracy as the clock star positions. As observations accumulate, the star positions will be steadily improved but, in the meantime, there is a risk of small systematic errors.

The Danjon transit instrument. An interesting attempt to modify the design of the transit instrument so as to secure a greater precision has been made by Danjon[12]. In this design the instrumental meridian is defined by purely optical methods. The pivots consequently serve merely as a support for the apparatus; they can therefore be turned on a lathe and small departures from cylindrical shape are immaterial.

The essential part of the instrument consists of an accurately worked Wollaston reversing prism, whose long face defines the instrumental meridian. The prism is placed in front of the objective of a telescope, so that it covers a central rectangular portion of the objective, leaving two parts, in the form of segments of a circle, uncovered. The prism and the telescope can be turned about an east-west axis, which passes centrally through the prism. When pointed at a star near meridian passage, two images are seen, one of which is formed by the light that has passed through the prism and the other by light that has passed through the two outer segments of the lens. The beams that produce the two images have the same principal ray. The two images move in opposite directions and at the moment of meridian transit (assuming that there are no errors of adjustment) the two images coincide. The direct image is observed before coincidence with an impersonal micrometer, as in an ordinary transit observation, and the reflected image is observed after coincidence over the same portion of the field. The mean of the two series is used to derive the instant of coincidence.

The long face of the prism is silvered and a collimating telescope is mounted on the axis of rotation, in an east-west direction. This telescope has a micrometer provided with two screws in directions at right angles. The coincidence of the wires with their images determines the normal to the mirror face of the prism. An essential part of the observation of each transit is the determination of this normal, the position of the instrument remaining unchanged between the transit and collimation observations. The azimuth of the autocollimating telescope is controlled by a distant azimuth mark.

The telescope is given a compact form by deflecting the light, after its passage through the objective, by a plane mirror, inclined at an angle of 15° to the optical axis, and then by an equilateral prism, placed on the axis of rotation, into the eye-piece, which is on the axis of rotation and therefore always in a fixed position. The reversing prism can be rotated through 180° so that the entry and exit faces can be reversed.

A small telescope of this design has been in use at the Strasbourg Observatory and has given excellent results. The time determined from equatorial stars agrees with that determined from zenithal stars to a few thousandths of a second; with transit instruments of conventional design, differences of a few hundredths of a second

frequently occur. This instrument deserves more extensive trials than it has yet received.

Equal-altitude instruments. The method of equal altitudes, as first suggested by Gauss in 1808, consisted in observing the times at which three stars reached the same altitude, this altitude not needing to be known accurately. If the positions of the three stars and the rate of the clock are known, the observations will give the clock correction or, in other words, the time, the latitude of the place of observation and the true altitude of the stars at the moments of observation. The method was devised by Gauss to eliminate instrumental errors of the sextant, which becomes merely an instrument of comparison. The first instrument designed specially for observations by the equal-altitude method was the *almucantar* of Chandler[13]. A rectangular trough contained mercury, supporting a rectangular float on which was mounted a telescope inclined to the vertical at an angle equal to the co-latitude. The trough could turn around a vertical pillar carrying the float with it. As the float turned, the end of the optical axis described a small horizontal circle (almucantar) passing through the pole. The telescope had a bent eyepiece and a series of horizontal threads, the transits of the stars across these threads being observed. Time, latitude and the positions of the stars are derived from observations that are all of the same type, whilst the flexure of the telescope is the same for all observations, as also is the mean refraction.

The advantages of the equal-altitude method have been utilized in several instruments, whose designs differ somewhat in detail. These instruments in portable form are well-adapted for field use and enable time to be determined with considerable precision. They are mostly based on the principle of utilizing two beams of light, one coming directly from the star and the other coming from the star but being reflected at a mercury surface. By means of an optical device, these two beams are directed into a telescope. As the star approaches a given altitude, which is determined by the optical device used, two images moving in opposite directions are observed in the telescope. The images coincide at the instant when the two beams of light entering the telescope become parallel, the star being then at the standard altitude.

In the *prismatic astrolabe* of Claude and Driencourt[14] the telescope is mounted horizontally and an equilateral prism is placed before the object glass, the face adjacent to the object glass being perpendicular to the optical axis. A mercury trough is placed slightly below and a little to the front of the prism. The telescope with the prism and mercury trough can be rotated about a vertical axis, the bearing being read from a horizontal circle. Rays of light from a star at an altitude of 60° enter the upper face of the prism directly and the lower face after reflection at the mercury horizon, both at normal incidence; after total reflections at the opposite faces the two beams pass normally through the third face into the telescope. The two images of the star are each formed by half the object glass and are slightly elongated vertically, being small spectra, the one inverted relatively to the other. The geometrical centre of the star is not quite coincident with the centre of the image, which introduces a magnitude equation. It is therefore desirable to observe stars

of nearly equal magnitudes. A change of the eyepiece setting affects the altitude at which coincidence of the two images occurs and, as the observation is very sensitive to a small change of focus, it is important not to touch the setting throughout a series of observations. A clamping screw is provided for the eyepiece, to prevent accidental shifts.

With a large prismatic astrolabe, 6 cm. aperture and 55 cm. focal length, time can be determined with an internal mean error of about 0·01 sec. The personal equation in observations with this instrument is large, as it may amount to one- or two-tenths of a second. There are several designs of apparatus for determining the personal equation of each observer.

The difficulties arising from personal equation can be avoided by making observations photographically. Errors arising from changes in eyepiece setting are eliminated at the same time. An adaptation of the prismatic astrolabe to photographic observations was made by R. Baillaud[15]. In his final model of the photographic instrument, the object glass had a diameter of 18 cm. and a focal length of 190 cm. As it is not possible to obtain a perfectly homogeneous prism of dimensions large enough to cover an objective of this size, the prism was reversed, its apex facing the object glass. The two inclined faces were silvered and the direct and reflected beams were reflected from these two faces. The photographic plate was placed in the focal plane, and the shutter was actuated by an electromagnet placed in the circuit of the clock contacts, so that it opened and closed alternately every full second, with an interruption at the 59th and 60th seconds. Two lines of equidistant images are produced on the photographic plate. The positions of corresponding points are measured under a microscope. The lag of the shutter is determined by recording the instants of its opening and closing on a chronograph, together with the clock signals. Stars to about the 5th magnitude can be observed; the number of stars is therefore limited. The observations are affected by magnitude equation, which can be eliminated by observing stars at both eastern and western passages. This requires the observations to extend over several hours. Some promising results have been obtained but the instrument has not been fully tried out by regular observations over a lengthy period.

The pentagonal prism astrolabe of Baker[16] has been designed to form two images by beams both of which are centred on the optical centre of the object glass, so that a change in focus does not affect the relative position of the images. The pentagonal prism has one right angle and four equal angles of $112\frac{1}{2}°$. It is placed with the right angle uppermost, so that the adjacent faces are inclined at angles of 45° to the horizontal. Stars are observed at an altitude of 45°, which permits a larger selection of stars than is possible with an equilateral prism. Light from a star at an altitude of 45° falls normally on one face and, after two internal reflections at faces that are silvered, emerges normally through the other 45° face. The central portion of the second face at which reflection occurs is left unsilvered. A small prism is cemented outside the lower face of such angle that light from the star, after reflection at the mercury horizon, enters this prism normally, passes through the unsilvered portion of the face of the pentagonal and emerges normally through

the same face as the direct beam. Thus the axis of each of the beams passes through the centre of the objective of the telescope, the axis of which is inclined to the horizontal at an angle of 45°.

In order to facilitate observation, the direct beam is split by a thin prism of angle about 1 min., covering half the outer zone of the object glass. Thus two direct images are obtained. These images and the reflected image cross the field of the telescope vertically and in opposite directions: coincidence occurs when the reflected image crosses the line passing through the two direct images, lying symmetrically between them.

Whereas in the original prismatic astrolabe only one observation of coincidence is possible, a multiplication of settings can be made with the Baker design. For this purpose, three thin prisms with angles in the ratio 1 : 2 : 4 can be inserted in the path of the rays that form the reflected image and each prism can be reversed. The prism of smallest angle is reversed each sight, that of the next prism every alternate sight and that of largest angle every fourth sight; thus eight observations of coincidence for any star can be obtained, employing deviations proportional to $+7, +5, +3, +1, -1, -3, -5, -7$.

The impersonal prismatic astrolabe of R. Baillaud[17] is a modification of the astrolabe of Claude and Driencourt by means of which the personal equations of the observers are to a large extent eliminated. Two small optically plane parallel plates of glass, of the same thickness and refractive index, are placed behind the objective (i.e. on the eyepiece side). The one covers the upper half of the objective, the other covers the lower half. Each plate can turn round a horizontal axis parallel to its own plane. By means of a screw, the two halves of which have right- and left-handed threads respectively, on which are threaded two nuts, the two plates are turned at a constant rate in opposite directions. The screw is driven by a motor and the rate of rotation can be adjusted to suit the azimuth of the star to be observed. The two images are set so that the line joining them is horizontal and they travel together along very close verticals. A differential motion is provided to correct the position of the plates, as necessary, so that the line joining the images remains horizontal. The mountings of the plates bear a series of electrical contacts, so that signals are sent to a chronograph, which records the times of passage of a star at a number of altitudes relative to known positions of the plates, which are the same for all the stars observed. In order to eliminate any personal equation in estimating the horizontality, observations to the east are always made with the rising image brought to the right of the descending image, whilst observations to the west are always made with the rising image to the left. Preliminary results with this instrument show considerable promise.

The circumzenithal instrument was designed by Nušl and Frič[18]. There have been several models but the general principle is the same in all of them. In the model that is used in the Czechoslovakian Geographical Service, two mirrors fixed relatively to one another at an angle of 50° are employed to throw two beams of light, one coming directly from the star and the other reflected from the mercury horizon, into the telescope. The mirrors are crossed and lie one on each side of the

vertical plane of symmetry of the telescope, so that the two beams cover the right and left halves of the object glass respectively. The mirror support can be turned by antagonistic screws through small angles in three mutually perpendicular planes, so as to obtain all the required adjustments. The two outer quarters of the object glass are covered by thin glass wedges of angle 50 and 90 sec., for the multiplication of sights. By using both coincidences and symmetrical groupings, seven sights can be obtained. In addition, a thin glass prism can be inserted in either direction in the path of the incident rays, enabling observations of the same star to be made at three altitudes, spaced 6 min. apart. In this instrument, the images from the two halves of the object glass are similar and elongated horizontally, which facilitates accurate setting, but there is some danger of the angle between the mirrors varying slightly. The best results with the instruments of this design give an internal probable error not much inferior to that given by a small transit instrument.

The bent telescope of de la Baume Pluvinel[19] is a modified form of almucantar instrument, with which observations can be made at the one particular altitude for which the telescope is designed. In de la Baume Pluvinel's telescope, this altitude is 60°. The telescope body consists of a rigid metal frame, whose upper surface carries the object glass on one side and the eyepiece, with a frame having ten horizontal wires in the focal plane, on the other side. The convergent beam from the object glass is reflected towards the eyepiece by a mercury horizon, supported from the base of the telescope body. The centre of the trough is on the optical axis and at a distance from the objective of half its focal length. The telescope can turn round a vertical axis, supported by a tripod with levelling screws and carrying an azimuth circle. The oblique transits of the star across the wire reticule are observed and it is possible to incorporate an impersonal micrometer to avoid personal equation.

The advantages of this instrument are that an error of tilt of the axis of rotation only introduces an error of second order in the observed altitudes, and refraction effects are small. The whole object glass is used so that the images are good. The observations are very sensitive, however, to small changes in the level of the mercury surface. A variation of a few microns can affect the instrumental altitude by as much as one second of arc. The mercury trough is of special design, with a very thin film of mercury, to reduce such variations to a minimum. In the world longitude programme of November 1933, a bent telescope of 95 cm. focal length was employed, along with other instruments, at the Paris Observatory and the mean error of a time determination is stated to have been about 0·01 sec.

Though the various types of equal-altitude instruments cannot yet be said to have given a greater precision in the determination of time than the small reversible transit instrument, the best of them are not greatly inferior in precision, and there are considerable advantages in the employment of a method that is fundamentally different from the meridian transit method. For field use, the equal-altitude instruments have the advantage of giving from the same observations a determination of the latitude, as well as of the time.

For the highest precision, a determination of the personal equations of the observers is essential. For a description of the various types of apparatus that are available for this purpose and for much other useful information on the equal-altitude method, reference may be made to an article by Mme E. Chandon and A. Gougenheim[20].

§ 3. STAR POSITIONS

The apparent sidereal time at which a star transits across the meridian is equal to the right ascension of the star. For the determination of time, the right ascensions of the stars observed must therefore be known. Conversely, when the sidereal time is known, the right ascension of a star can be found by observing the time of its meridian transit. The determination of time is normally based upon the observations of the brighter stars, which have been under observation for many years. By repeated successive approximations, the accuracy of the positions has been gradually improved. On any one night, assuming the rate of the clock to be known from observations of the same stars on consecutive nights, the observations will give differences of right ascension between pairs of stars. These must be carried right round the sky until the first star is observed again. The closing error of the completed network must be averaged over the individual stars. Thus a series of differencies of right ascension is obtained. The position of the equinox, or First Point of Aries, which is adopted as the starting point for the measurement of right ascensions, is determined from observations of the sun, Mercury and Venus.

In this way, with any given instrument, a series of star-places can be derived. By comparing observations at different epochs, the proper motions of the stars can be deduced. But, in spite of all the care taken in the observations, it is found that the results obtained with different instruments show differences that, apart from accidental errors of observation, are of a systematic character. Within a given zone of declination, the right ascension differences will vary with right ascension; within any hour of right ascension, the differences will vary with declination. Thus the instrumental network of star positions is a somewhat distorted representation of the actual network.

A fundamental system of star-places is derived from the comparison of the results provided by a large number of instruments, distributed over the world, and from their combination in such a way that the errors inherent in the separate instruments are largely averaged out. A fundamental system provides, moreover, what no single instrument can give, a network of stars covering the whole sky.

The principal fundamental systems that have been used of recent years are those of Auwers (revised)[21], Newcomb[22], Boss[23] and Eichelberger[24]. The Eichelberger system has been used in the *Nautical Almanac*, the *American Ephemeris* and the *Connaissance des Temps* since 1928; previously these ephemerides used either the Newcomb or Boss positions. The revised Auwers system is used in the *Berliner Jahrbuch*. Of these fundamental systems, that of Eichelberger gives the most correct representation of the star-places for the epoch of 1925. This system was based largely on Washington and Cape Catalogues of recent date. But the system

of the proper-motions is seriously in error, so that with passage of time the star-places on this system would become increasingly inaccurate. The revised Auwers system provides a much better representation of the proper-motions than the Eichelberger system but, partly because of the time that has lapsed since it was derived, the representation of the star-places for the epoch of 1925 is not so satisfactory as on the Eichelberger system.

A further revision of the Auwers system has recently been completed at the Rechen Institut in Berlin. A very exhaustive discussion of all the available material has been given and the new fundamental catalogue (*Dritter Fundamentalkatalog des Berliner Astronomischen Jahrbuchs*, denoted by FK3) undoubtedly provides the best representation both of the positions and of the proper-motions of the stars that is possible at the present time. It will be adopted in due course in all the national ephemerides and is likely to remain for many years the accepted fundamental system of star-places.

The FK3 system of star-places shows differences from those of the revised Auwers system, which vary both with right ascension and with declination[25]. There is, for instance, at the equator a range in the smoothed hourly means of 0·004 sec. At declination $-50°$, the range is 0·011 sec. There are also systematic differences, varying both with right ascension and declination, in the proper-motions. The other fundamental systems also show systematic differences between each other. The corrections to the places of individual stars may differ considerably from the mean corrections, e.g. the FK3 system indicates corrections as shown in table 6 to the revised Auwers places of three stars[26]:

Table 6

Star	Right ascension	Declination	Correction (1925·0)
α Puppis	$7^h\ 50^m$	$-40°\ 24'$	$-0·051$ sec.
ζ Puppis	$8^h\ 1^m$	$-39°\ 49'$	$-\ ·007$ sec.
q Puppis	$8^h\ 16^m$	$-36°\ 27'$	$+\ ·020$ sec.

Assuming the new places to be correct, the errors of the places at present given in the *Berliner Jahrbuch* show a range of 0·07 sec. These errors in star-places must affect the accuracy with which time can be determined. The errors that are periodic in right ascension would cause an apparent annual variation in longitude between two observatories, if one observatory reduced its time observations on the basis of one system of star-places whilst the other observatory used a different system.

The *Nautical Almanac*, and other ephemerides, give the mean places of the stars for the beginning of the year and the apparent places for the upper transit at Greenwich for 10-day intervals or, in the case of close circumpolar stars, for each day. The mean place for the beginning of the year is the position of the star referred to the mean equinox at the beginning of the Besselian fictitious year. The apparent place is the position referred to the true equator and equinox of date and is corrected for precession and nutation, and also for stellar aberration. It is therefore the position in which the star is seen by the observer.

For the reduction of time observations, the observed place must be compared with the apparent place for the time of observation. At present none of the principal ephemerides contains the apparent positions of sufficient stars to meet all the requirements of the time observer, with the result that the apparent places of some stars have to be extracted from one almanac, those of others from a second almanac and so on. In accordance with a resolution of the International Astronomical Union[27] a separate annual volume of apparent places of 1535 stars will be published, commencing with the year 1941; the number of stars in any one of the ephemerides, for which apparent places are given, will be reduced to about 200. The computation of the apparent places of these stars will be shared by the almanac offices in Berlin, Paris, San Fernando and Washington whilst the British Nautical Almanac Office will be responsible for the printing and publishing of the volume. The new volume of apparent places will provide the apparent places on the system FK 3 of sufficient stars to meet the requirements of observers in any part of the world who desire to determine time or longitude.

§ 4. STANDARD CLOCKS

Pendulum clocks. Determinations of time by astronomical observations require a clear sky, and intervals of several days may elapse without observations being possible. For the daily distribution of time by means of various types of time signals it is necessary to rely upon one or more precision timekeepers. The accuracy of the time distributed by an observatory is therefore conditioned by the irregularities in the rates of the standard timekeepers. Within the last two decades or so there has been a considerable improvement in the performance of precision timekeepers.

The Riefler pendulum clock marked the first great advance in reliability of performance[28]. In the Riefler clock, the pendulum is impulsed by the bending of the suspension spring at the moment when the pendulum passes through its middle position. The only connexion between the pendulum and the clock train is through the suspension spring; the pendulum swings entirely freely except at the moment of impulsing. The flexing of the suspension spring is produced by a special escapement, the anchor of which is carried on the pendulum. The precision Riefler clock is provided with automatic electric winding at intervals of about 40 sec., has an invar rod and is enclosed in an airtight case, to eliminate barometric effects. The air pressure within the case is normally about 26 inches, and the rate of the clock can be changed, without unsealing the case, by varying the pressure. A change of pressure of 1 mm. produces a change in rate of 0·018 sec. per day. The rate of a good Riefler clock, when mounted in a room where temperature is maintained constant, can be relied upon to an accuracy of about 0·01 or 0·02 sec. per day.

A more complete separation of the pendulum from the clock train is achieved in the Shortt free pendulum clock. The clock consists of two parts: a master pendulum mounted in an airtight case, swinging freely except at the moments of

impulsing, and an auxiliary or "slave" clock, for recording and distributing the time, which is kept in synchronization with the master pendulum. Several horologists had previously attempted to construct a clock in which all the work was put upon a slave clock; in particular, Gill at the Cape Observatory had constructed a clock on these principles, in collaboration with Horace Darwin, which had performed well for short periods of time, but the clock was so complicated and contained so many contacts that it was found impossible to keep it running continuously. To Shortt must therefore be given the credit for the first successful design of such a clock.

The principle of the Shortt clock is briefly as follows: the master pendulum swings freely in an airtight case, at a pressure of about $\frac{3}{4}$ inch of mercury, except when it is impulsed. The impulse is given once every half minute. The pendulum carries a small wheel. The impulse arm (released by an electric current sent by the slave clock) falls on the top of this wheel as the pendulum is passing through its middle position. The pallet on the impulse arm rests on this wheel and, as the pendulum swings out, the pallet travels down the edge of the wheel, finally falling clear. It then trips an arm, which falls, making an electric contact and sending a current to the synchronizing arrangement on the slave clock. The slave clock is a standard synchronome electric clock, provided with seconds and half-minute circuits. This clock is rated so that, when running as an independent clock, it loses about 6 sec. per day. The synchronizing current from the free pendulum passes round the coils of an electromagnet adjacent to the slave pendulum; the armature of this electromagnet carries a tongue, which can engage the bent top of a long leaf spring carried on the slave pendulum. When this happens, the end of the spring is held while the pendulum swings until, on the return swing, the spring is carried away. The spring is flexed as the pendulum swings, and the pendulum is accelerated by $\frac{1}{10}$th of a second. The natural losing rate of 6 sec. per day corresponds to $\frac{1}{10}$th of a second in half a minute, so that in order to keep the slave pendulum in time with the master pendulum it requires acceleration about once per minute. Suppose the proper phase relation between the two pendulums has been established, and that the slave pendulum has just received an acceleration. Half a minute later the synchronizing tongue will again fall, but the slave pendulum will not have dropped behind sufficiently for the tongue to engage the spring; it will merely fall on top of it. At the end of the next half minute, the slave pendulum will have dropped further behind and the spring will again be caught by the synchronizing arm. The release of the free pendulum impulse arm is produced by the current from the half-minute circuit of the slave clock and, since there is a practically constant phase relation between the master and slave pendulums, the fall of the impulse arm always occurs when the free pendulum is in practically the same position. As the arm falls on the top of the small wheel on the free pendulum, the slight variations that may occur on account of the possible variation of 0·004 sec. in the phases of the two pendulums are immaterial.

The first Shortt clock was installed at the Royal Observatory, Edinburgh, in 1921 and reports on its performance have been given by Sampson[29]. More ex-

haustive discussions of the performance of the Shortt clocks subsequently installed at the Royal Observatory, Greenwich, have been made by Jackson and Bowyer[30]. These investigations have shown that the performance of this type of clock is superior to that of any other design of pendulum clock. The clock is more accurate than the time determinations so that, if the apparent errors of the clock as indicated by the transit observations are plotted, the best representation of the clock errors is obtained by drawing a smooth curve between the plotted points. At the Greenwich Observatory the observations are referred to a *mean clock*, which represents the mean of five Shortt clocks. Time signals are based on extrapolation from the smooth curve representing the error of the mean clock.

If the rate of the clock were absolutely uniform the clock error could be inferred from a straight line drawn through the clock errors determined by observation. Though the Shortt clock has the highest precision yet attained with a pendulum clock, it is not free from sudden changes of rate and small irregular fluctuations in rate. These are of the general order of a few milliseconds a day, though the sudden changes of rate may be considerably larger. The causes of these changes are obscure. Small variations in the energy imparted to the pendulum when impulsed can occur through the pallet of the impulse arm falling clear of the impulse wheel at slightly different points, if the wheel is not perfectly true or if the bearings are slightly worn. The semi-amplitude of swing (about 50 min.) is increased on impulsing by 6 sec. A change in semi-amplitude of 2 sec. will alter the rate by 0·0014 sec. a day or by about 0·5 sec. in a year. Some of the variations of rate are accompanied by change of arc and Jackson and Bowyer have shown that if corrections are applied to reduce to a constant arc the performance of the clock is improved. It is possible that automatic control of the arc by photo-electric means would result in an improved performance. This is being done by Loomis[31], but no detailed results have yet been published. It is suspected at Greenwich that some of the changes of rate are due to changes in the suspension spring; the elinvar suspension springs supplied with the clocks have been replaced, on account of the somewhat capricious behaviour of elinvar, by steel suspension springs. When the temperature of the clock room is well controlled there is nothing to be gained by the use of elinvar.

Small irregular fluctuations of the order of 0·01 or 0·02 sec. cannot be determined accurately by observations that may be in error by as much as 0·04 or 0·05 sec. and which are spaced irregularly, with frequent gaps of several days. There seems little hope at present of any marked improvement in the accuracy of the time determinations themselves. There is more hope of controlling the performance of the pendulum clock by some form of vibration clock.

The pioneer work in quartz crystal oscillators was carried out at the Bell Telephone Laboratories in New York under W. A. Marrison[32]. The crystals oscillate 100,000 times a second and they can be compared with each other with such precision that a change of 1 part in 10^{10} can be detected over a period of 10 sec. Since 1929 Loomis has regularly compared these crystals with his Shortt clocks at Tuxedo by means of a private wire between New York and Tuxedo

and a special form of chronograph. His comparisons were used by Brown and Brouwer[33] to detect the effect of the gravitational attraction of the moon on the clock pendulum, an effect which has a semi-amplitude of 0·00015 sec. and a period of half a lunar day.

The best results yet obtained in the use of quartz crystal oscillators as standards of frequency or time have been by Scheibe and Adelsberger at the Reichanstalt[34]. The later crystals are short rods of rectangular cross-section with approximate dimensions of 91 mm. long and 11·4 mm. deep. The long axis is in the direction of the electrical axis of the crystal, the sides of the cross-section being parallel respectively to the neutral and the optical axes. The mode of vibration is longitudinal along the long axis, the second harmonic, with a frequency of 60,000 c. per sec., being used. The crystal is supported at the nodes in a glass cylinder containing hydrogen at low pressure. The temperature is maintained constant at 35·5° C., at which temperature the frequency is practically independent of small temperature variations. The frequency is divided in three stages: 60,000 to 10,000; 10,000 to 1000; and 1000 to 500 and 250. The vibrations of lowest frequency are recorded by a phonic motor.

An account of the performance of two such clocks at the Geodetic Institute, Potsdam, over a period of more than one year has been given by Pavel and Uhink[35]. Their discussion shows that the irregularities of rate of these clocks are less than 0·001 sec. per day, so that it may be said that they are capable of keeping a constant rate with something like 10 times the precision of the best pendulum clocks. Pavel and Uhink claim to have detected variations in the rate of rotation of the earth during the period covered by the observations; but this claim does not seem to be justified, as the clocks may merely be revealing systematic errors, partly of a seasonal nature, in the time observations.

A quartz crystal oscillator of a different form was developed at the National Physical Laboratory by the late D. W. Dye, whose work has been continued by L. Essen[36]. The crystal is in the form of an annular ring, cut in a plane perpendicular to the optic axis. The internal and external diameters are 75·3 and 98·6 mm. respectively, the thickness being 16·3 mm. The mode of vibration is one of uniform circumferential strain, in which the ring expands and contracts radially as a whole, the frequency being 20,000 c. per sec. The dimensions of the ring were chosen so that there should be no other mode of vibration present. To preserve a frequency stability of 1 part in 10^8, the temperature is maintained constant to 0·003° C. The frequency is divided in two stages to 1000 c. per sec., the output being amplified to drive a phonic motor, which provides contacts at 1·0, 0·1 and 0·01 sec. A comparison between two such oscillators revealed maximum fluctuations of ± 5 parts in 10^9 over a period of 24 hr. or $\pm 0{\cdot}0005$ sec. per day.

With this mode of vibration there are no nodes, which causes some difficulty in supporting the crystal. A ring oscillator of another type has therefore been developed at the National Physical Laboratory, with a frequency of 100,000 c. per sec.[37]. An overtone mode of vibration is used, permitting the ring to be held rigidly between the electrodes at three nodes. The electrode system is simpler and the

vibrating ring does no work on the surrounding atmosphere, so that an evacuated enclosure is not essential. A quartz oscillator clock of this type is under construction at the Laboratory for the Royal Observatory, Greenwich, and will be used to control the irregular fluctuations and the changes in rate of the pendulum clocks.

A vibration clock that embodies a different principle has been developed at the National Physical Laboratory by J. E. Seares and G. A. Tomlinson[38]. It consists of an elinvar bar, one metre in length, maintained in vibration by electrostatic forces applied at one end. The bar is housed in an exhausted steel vessel, surrounded by a water bath, controlled in temperature to within 0·002° C. The frequency is approximately 2366 c./sec. and is applied direct to a phonic motor that has 91 poles, the rotor driving through a 26 : 1 reduction gear, giving the first spindle a speed of approximately one revolution per second. This clock has an accuracy comparable with that of the quartz oscillator clock. The additive irregularities in the two clocks rarely exceed 0·001 sec. in a prolonged trial[39].

Vibration clocks of these types must be regarded as still in the development stage. It is difficult to keep the electrical conditions sufficiently invariable to maintain stability of rate over long periods. Stoppages are also comparatively frequent and it is not generally possible to obtain a long continuous run. The temperature and pressure require to be much more closely controlled than in the case of the pendulum clock. Though it is probable that vibration clocks will eventually become the standard timekeepers in observatories that are concerned with the determination and distribution of time, their main role at present is the accurate subdivision of time between transit observations and the control of the small irregular fluctuations of the pendulum clocks.

A less accurate type of vibration clock, which can be kept running continuously without difficulty, may be mentioned. This is a valve maintained elinvar tuning fork, vibrating with a frequency of 1000 c./sec., developed by Dye at the National Physical Laboratory[40]. For a frequency stability of 1 part in 10^7, the temperature must be maintained constant to 0·006° C. This fork is capable of maintaining a constant rate for several weeks at a time to within 2 or 3 parts in 10^7. Variations of frequency are due to changes of temperature, pressure or voltage supplies. The fork enclosure was not originally designed for maintaining a constant pressure and is not sufficiently rigid so that, although it is airtight, pressure variations of 2 or 3 mm. occur. It is believed that by the complete elimination of voltage and pressure changes a frequency stability over short periods of a few parts in 10^9 and over long periods of 1 part in 10^7 would be obtained. This would be an appreciable improvement on the best pendulum clocks.

§ 5. THE VARIABILITY OF THE UNIT OF TIME

The rotation of the earth provides the standard clock, to which all other measures of time are referred for comparison. It has been tacitly assumed in the present article that the unit of time, whether we adopt the mean solar or sidereal day or second, is constant and does not vary. Within recent years it has been

established that the length of the day is not constant and it remains to discuss its variations and the possibility of detecting and controlling them by means of mechanical clocks.

Slight variations in the length of the day should be shown most easily by observation of the position of the moon, because of the moon's rapid motion in longitude. The variations in the length of the day would cause apparent variations in the mean longitude of the moon, which should be revealed by comparison between the observed positions of the moon and those computed from pure dynamical theory. Such differences between observation and theory have been found to exist and are of two types, which will be considered separately.

(i) *The secular acceleration of the moon's mean longitude.* The mean longitude of the moon contains a term of the form $\frac{1}{2}aT^2$, where T denotes the time. This term implies that the mean motion of the moon in longitude is not constant but is gradually changing. When T is expressed in centuries, the quantity a is called the *secular acceleration* of the moon's motion. Its existence was first discovered by Halley, who found that it was required in order to represent by one formula observations of eclipses by the Babylonians and by the Arabian astronomer Albategnius in the ninth century A.D. and the current observations of the moon's position. In the middle of the eighteenth century Tobias Mayer suggested that this secular acceleration might be the result of a slowing down of the earth's rotation through the friction of the tides. But in 1787 Laplace found that the gravitational action of the planets on the moon would produce a secular acceleration and, according to his computations, the acceleration due to this gravitational action was in close agreement with the observed acceleration. Adams, however, in 1853 revised Laplace's computations and carried them to a further order of approximation and found that the theoretical acceleration was only about one-half of the observed acceleration. Fotheringham's discussions of ancient records of solar and lunar eclipses and of observations of occultations, and of Hipparchus's determinations of the equinox led to a much improved determination of the moon's secular acceleration and confirmed the discordance[41].

(2) *Irregular fluctuations of the moon's mean longitude.* It was discovered by Newcomb, in his researches on the motion of the moon, that the moon showed irregular deviations from its computed path[42]. The chief fluctuation appeared to be well represented by a harmonic term with an amplitude of about 14 sec. and a period of about 270 years. This has become known as the "great empirical term" and, though it could not be accounted for by theory, it was incorporated by Brown in his "Tables of the Motion of the Moon". Superimposed on this major fluctuation were minor fluctuations, with no regular period, amounting to three or four seconds of arc.

If the irregular fluctuations and the unexplained portion of the secular acceleration of the moon's longitude are produced by changes in the rate of rotation of the earth, it should be possible to find corroborative evidence from the motions of Mercury, Venus and the sun, the magnitudes of the effects on these bodies being smaller because of their smaller mean motions. A secular acceleration in the

longitude of the sun was detected by Cowell in 1906[43] and Fotheringham's investigations[41] gave an improved value. Irregular fluctuations exist also in the longitudes of these bodies. For the sun and Venus, the study of these fluctuations must be restricted to the period subsequent to 1830, as the meridian observations before that date were too erratic to provide any information of value. For Mercury however, the observations of transits of the planet across the sun, back to the year 1677, can be used. These transits were discussed by Innes[44]. It was shown by de Sitter and by Spencer Jones[45] that the fluctuations in the longitudes of the sun, Mercury and Venus closely reproduced those in the longitude of the moon, when allowance was made for the different mean motions. Thus, whenever the moon is losing time (falling behind its computed position), the sun and planets are also losing time, and whenever the moon is gaining time, the sun and planets are also gaining time. The only way in which these results can be explained is on the assumption that the earth is not a perfect clock.

The portions of the secular accelerations that remain unaccounted for by pure dynamical theory can be explained by a retardation of the rotation of the earth as a result of tidal friction, particularly in narrow straits and shallow seas such as the English Channel, the Straits of Malacca and the Bering Sea. The calculations of Taylor and of Jeffreys[46] have shown that the energy dissipated in this way is of the right order of magnitude. The irregular fluctuations cannot be explained by tidal friction, as this can only cause a retardation in the rotation, whereas the fluctuations correspond to both retardations and accelerations. They must be associated with changes in the moment of inertia of the earth. Brown[47] has suggested vertical oscillations of the crust of the earth, the source of the oscillations being due to oscillatory changes taking place in the isostatic layer or below it. A uniform shrinkage or expansion of the whole earth, involving a change of radius of a few inches, would be sufficient to account for the observed variations. De Sitter has pointed out that a very slight adjustment of the shape of the layers of equal density in the inside of the earth, of which the effect on the dimensions of the outer surface need not exceed a fraction of a millimetre, could equally well occur for the observations.

The two effects, tidal friction and changes of moment of inertia, are reflected in the apparent positions of the moon and of the sun and planets in different ways. Tidal friction on the earth influences the orbit of the moon, because the tidal wave is caused by the attraction of the moon and it must in turn attract the moon, creating a tendency to accelerate its motion. This causes the moon to recede from the earth and consequently its period of revolution is increased and the moon will be apparently retarded on this account. Tidal friction has, however, no direct action on the sun or planets. Change of moment of inertia of the earth, on the other hand, has no effect on the orbit of the moon. The departures of the motions of the moon, sun, Mercury and Venus from dynamical theory can be satisfactorily accounted for by a combination of the two effects.

According to the results of de Sitter, the average increase produced by tidal friction in the length of the day over the last 25 centuries is +0·0022 sec. per

century, but since 1870 it has averaged +0·0037 sec. per century. The cumulative effect in the course of a year amounts to 0·014 sec., which is far outside the limits of detection by the best vibration clocks.

The effects of changes of inertia are much larger. Such changes seem to occur almost discontinuously, though the observations are not of sufficient precision to distinguish between a sudden change and a change that is progressive through two or three years. The adopted excesses of the length of the day over its average value, according to de Sitter, are given in table 7.

Table 7

Date	Excess (sec.)
From 1664 to 1755	+0·00134
,, 1755 ,, 1786	+0·00089
,, 1786 ,, 1864	−0·00091
,, 1864 ,, 1876	−0·00357
,, 1876 ,, 1897	−0·00186
,, 1897 ,, 1918	+0·00153
,, 1918 to date	−0·00186

The greatest changes occurred in or near 1897 and 1918, amounting in each case to a change in the length of the day of 0·00339 sec., amounting in a year to 1·24 sec. There would seem to be some possibility of detecting such a change by a vibration clock, though it is just beyond the present capabilities of the best pendulum clocks. There is, moreover, the possibility that the readjustment of masses in the interior of the earth, which alters the moment of inertia, may have an effect on the earth's gravitation. Clocks of the vibration type may therefore be more suitable than pendulum clocks for detecting such changes in the moment of inertia.

To detect a change in the length of the day it will be necessary for the vibration clock to maintain its rate steadily within a fraction of a second a year and to keep on doing this for a number of years. If an observatory is provided with three such clocks, the effect of a temporary stoppage of any one clock could be allowed for because comparisons of relative rates between this clock and the other two clocks before and after the stoppage, which can be made to a very high degree of accuracy, would enable any discontinuity of rate resulting from this stoppage to be allowed for.

A change in the length of the day, of the order of 0·003 sec., if it occurred instantaneously should be detectable because the rates of all vibration clocks would be changed by this amount.

There has been no sudden change in the length of the day since about 1918. The conclusion from this discussion is that, when such a change next occurs, there is a good chance that it will be detected by vibration clocks, provided that the change is comparable with those that occurred in or near the years 1864, 1876, 1897 and 1918.

REFERENCES

(1) MASKELYNE. *Astronomical Observations made at the Royal Observatory, Greenwich, from M.DCC.LXXXVII to M.DCC.XCVIII*, p. 339 (1799).
(2) *Greenwich Observations*, 1912, p. A xvii.
(3) GILL. *Mem. R. astr. Soc.* **46**, 1–172 (1881) [vide Chapters IX and X].
(4) BOSS. *Preliminary General Catalogue*, Carnegie Inst. Washington, No. 115 (1910).
ROY. *Systematic Corrections and Weights of Catalogues*, Dudley Observatory (1921).
(5) REPSOLD. *Astr. Nachr.* **123**, 177 (1889).
(6) REPSOLD. *Astr. Nachr.* **141**, 279 (1896).
(7) STRUVE. *Astr. Nachr.* **155**, 353 (1901).
See also COHN, *Astr. Nachr.* **157**, 357 (1902).
(8) GILL. *History and Description of the Royal Observatory, Cape of Good Hope*, pp. 51–58 (1913).
(9) *Cape Annals*, **13**, pt. 2, xvii (1927).
(10) ROSS. *Special Pub. U.S. Coast and Geodetic Survey*, No. 27 (1915).
(11) LITTELL and WILLIS. *Astr. J.* **40**, 7 (1930).
LITTELL. *Astr. J.* **43**, 9 (1933).
(12) DANJON. *Bull. Astr.* Ser. II, **9**, 131 (1934).
(13) CHANDLER. *Harvard Annals*, **17**, 1–222 (1887).
(14) CLAUDE and DRIENCOURT. *Rev. gén. Sci. pures et appliquées* **16**, 1071 (1905); *J. Phys.* **6**, *4th series*, 950 (1907). *Description et usage de l'astrolabe à prisme* (Gauthier-Villars, Paris, 1910).
(15) BAILLAUD, R. *Rev. Opt.* (*théor. instrum.*) **3**, 71 (1924).
(16) BAKER. *Geogr. J.* **77**, No. 5 (1931).
(17) BAILLAUD, R. *Bull. géod. int.* No. 41, 1 (1934). *Ann. Obs. Besançon*, Nouvelle Série, **2**, 13 (1936).
(18) NUŠL and FRIČ. *Pub. Obs. Nat. Prague*, Nos. 1 and 2, 1925.
See also *Bull. Acad. Sci. Bohême*, 1901, 1903, 1906, 1908.
(19) DE LA BAUME PLUVINEL. *C.R. Acad. Sci., Paris*, **174**, 535 (1922). *Bull. Soc. Astr. Fr.* **38**, 60 (1924).
(20) CHANDON and GOUGENHEIM. *Hydrogr. Rev.* **12**, No. 1 (1935).
(21) *Neuer Fundamentalkatalog des Berliner Astr. Jahrbuchs, Veröff. Astr. Rechen. Inst.* No. 33.
(22) *Catalogue of Fundamental Stars. Astr. Papers Amer. Ephemeris*, **8**, pt. 2, 1899.
(23) *Preliminary General Catalogue*, Carnegie Inst. Washington, 1910.
(24) *Positions and Proper Motions of* 1504 *Standard Stars for the Equinox* 1925·0. *Astr. Papers Amer. Ephemeris*, **10**, pt. 1, 1925.
(25) NOWACKI. *Astr. Nachr.* **255**, 301 (1935).
(26) *Reduktion des NFK auf den FK*3 *für* 1925·0 *und für* 1934·5, 1935·5, 1936·5. (*Anhang zum Berliner Astronomischen Jahrbuch* 1936, Berlin, 1934.)
(27) *Trans. inst. astr. Un.* **5**, 370 (1935).
(28) *Präzisions-Pendeluhren und Zeitdienstanlagen für Sternwarten.*
S. RIEFLER. (Th. Ackermann, München, 1907.)
(29) SAMPSON. *Proc. roy Soc. Edinb.* **44**, 56 (1924); **48**, 161 (1928).
(30) JACKSON and BOWYER. *Mon. Not. R. astr. Soc.* **88**, 465 (1928); **89**, 239 (1929); **90**, 268 (1930); **91**, 291 (1931).
(31) LOOMIS. *Mon. Not. R. astr. Soc.* **91**, 569 (1931).
(32) MARRISON. *Bell Telephone Lab. J.* **8**, 493 (1929).
(33) BROWN and BROUWER. *Mon. Not. R. astr. Soc.* **91**, 575 (1931).
(34) SCHEIBE and ADELSBERGER. *Hochfrequenztech. u. Elektroakust.* **43**, 47 (1934); *Phys. Z.* **33**, 835 (1932).
(35) PAVEL and UHINK: *Astr. Nachr.* **257**, 365 (1935).
(36) ESSEN. *Proc. roy. Soc.* A, **155**, 498 (1936).
(37) *N.P.L. Report for the Year* 1934, p. 61 (1935).

(38) TOMLINSON. *Observatory*, **57**, 191 (1934).
(39) *N.P.L. Report for the Year* 1934, p. 131 (1935).
(40) DYE and ESSEN. *Proc. roy. Soc.* A, **143**, 285 (1934).
(41) FOTHERINGHAM. *Mon. Not. R. astr. Soc.* **69**, 204, 666 (1909); **75**, 377, 395 (1915); **78**, 406 (1918); **80**, 578 (1920); **81**, 104 (1920).
(42) NEWCOMB. *Washington Observations*, App. 2, 1875. *Amer. Ephemeris Papers*, 9, pt. 1 (1912). See also *Mon. Not. R. astr. Soc.* **63**, 316 (1903); **69**, 164 (1909).
(43) COWELL. *Mon. Not. R. astr. Soc.* **66**, 3, 352 (1906).
(44) INNES. *Union Obs. Circ.* No. 65 (1925).
(45) DE SITTER. *B.A.N.* **4**, 21 (1927).
SPENCER JONES. *Cape Annals*, **13**, pt. 3, 30 (1932).
(46) TAYLOR. *Philos. Trans.* A, **220**, 1 (1919).
JEFFREYS. *Philos. Trans.* A, **221**, 239 (1920).
(47) BROWN. *Yale Univ. Obs. Trans.* **3**, 209 (1926).

THE ADSORPTION OF GASES BY SOLIDS

By ERIC K. RIDEAL, F.R.S.

In recent years more attention has been given to the kinetics of adsorption processes. This has been due to the fact that whilst two distinct types of adsorbed states are recognized, i.e. the Van der Waals' or physical, and chemiadsorbed states respectively, the slow rate of attainment of equilibrium observed in many cases has been attributed both to (1) an appreciable energy of activation involved in the chemisorptive process—activated adsorption[1], and to (2) diffusive or migratory processes involving an energy of activation[2].

The realization that chemiadsorbed films on metals are difficult to remove and that when once removed great precautions have to be taken to maintain their cleanliness, has resulted in the development of several new techniques in which measurements are made on a single wire as adsorbent surface. Analysis of surface films thermally by observations both on the accommodation coefficient and by resistance changes have been made, especially by Roberts[3], thermionically by Langmuir[4], Becker[5], and de Boer[6], photoelectrically by Bosworth[7], and by means of the contact potential by Bosworth and Rideal[8].

A series of papers on the accommodation coefficient have been published by Jackson and Mott[9], Jackson and Howarth[10], Landon[11], and Lennard-Jones and his co-workers.

The Langmuir isotherm in its simplest form is expressed by the equation

$$x = \frac{ap}{1 + ap}.$$

This has been derived thermodynamically by Volmer[12] and statistically by Fowler[13]. We note that the parameter a is in fact proportional to the product of two quantities, the condensation coefficient c and τ the lifetime of the adsorbed molecule on the surface. Lennard-Jones, Strachan and Devonshire[14] have derived explicit formulae for c and τ. Lennard-Jones and Devonshire[15] have recently examined the problem of diffraction of molecules which are running along a surface as well as the inverse process of capture by a surface of a molecule incident on the surface at a small grazing angle; they find it possible to develop expressions for the excited states of an adsorbed atom and to develop the Langmuir isotherm.

The equation given above refers to the adsorption of a gas by one point of contact, i.e. not involving surface dissociation. Further, it is assumed that there is no interaction between the adsorbed particles, and the equation is generally applicable to those cases where weak fields are involved, e.g. for gases attracted by Van der Waals' forces to surfaces already saturated by a chemiadsorbed layer. In general for chemiadsorbed gases there exists an interaction energy due to dispersive forces. These dispersive forces have been treated in two essentially different ways,

depending on whether short- or long-range forces need to be considered. If only short-range forces exist, the energy is assumed to vary only with the number of nearest neighbours and is in the form:

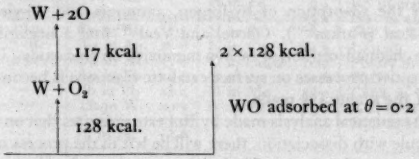
$$x = x_0 - fx,$$

where x is the number of pairs of nearest neighbours. This treatment has been developed by statistical methods by Fowler[16] and Peierls[17], and kinetically by Roberts[18]. At the same time, however, it must be recognized that all of these chemisorbed films possess a vertically orientated dipole, and so there exists an electrical interaction which varies relatively slowly with the distance apart of the interacting atoms and which it is quite impossible to treat by considering nearest neighbours only. For films of hydrogen on tungsten this electrical interaction accounts for only some 4 per cent of the total interaction energy, and it is clear therefore that for all such films the methods of Fowler and Peierls are available. For films of sodium or potassium on tungsten, however, Bosworth has shown that the electrical interaction accounts for some 90 per cent of the total interaction, and so for such films the method of nearest neighbours is inadequate, and one must then use an equation derived by Topping[19] which gives the spreading force in terms of dipole interaction and ignores short-range forces. Finally it may be noted that these dipoles are polarizable, and consequently the dispersive forces due to dipole interaction may vary less rapidly than that calculated by Topping's equation. It is clear, however, that in a perfectly general treatment of interaction between adatoms both long-range and short-range forces should be considered. The mathematical difficulties in the calculation both of the most probable position of the occupied lattice points and of the mutual potential energy of those points are great, and to date no calculation of surface-spreading forces applicable to a generalized gas-solid system has been made. From the experimental side, spreading forces (F) have been calculated from the rate of evaporation for films of caesium by Langmuir and sodium by Bosworth, and empirical F, A curves constructed.

Another factor implicit in the Langmuir equation in its simplest form is that the condensation coefficient is independent of the surface density of the adsorbed gas. This, as we shall see, is clearly not the case for a gas which undergoes dissociation on adsorption into two or more adsorbed fragments but is also not the case for many simple monatomic systems. Thus for caesium on tungsten and oxygen on cuprous oxide, to cite two dissimilar cases, condensation on the solid surface occurs not only by impact from the gas phase on the bare surface, but also by a migratory process from atoms or molecules adsorbed on the top of the primary layer. Condensation thus takes place much more rapidly than that calculated (with the aid of the Herz-Knudsen equation) for filling up the empty spaces.

One of the processes of chemisorption of most frequent occurrence is that in which molecular dissociation on adsorption is involved, resulting in the formation of a surface complex such as W-O, W-N, W-H. The isotherms have been developed by Fowler and Lennard-Jones. It is by no means a simple matter to prove that such disruptive actions take place in the act of chemisorption. Examination of the

nature of the evaporated gas is not conclusive unless the compound evaporates, and frequently it is only a compound more saturated than that which exists on the surface which is actually collected. The following diagram[20] demonstrates how difficult it is to discriminate between the evaporation of atoms and of molecules from purely thermal measurements.



The measured heat of desorption of oxygen from tungsten (for θ about 0·2) is 128 kcal. If evaporation takes place in the form of molecular oxygen 128 kcal. are required to evaporate 32 g., but if atomic oxygen evaporates, 128 kcal. will evaporate only 16 g., 256 kcal. being required for 32 g. of O_2. To dissociate the 32 g. of molecular oxygen, a further 117 kcal. are required, yielding a total energy of 245 kcal. required for molecular evaporation followed by dissociation. It is therefore difficult to decide from measurements of the heat of desorption in what form the oxygen evaporates.

Direct measurements by Roberts gave a value of $0{\cdot}7 \times 10^{15}$ molecules of hydrogen and of oxygen per cm^2 of tungsten. Since the crystal lattice constants indicate that there are $1{\cdot}4 \times 10^{15}$ atoms of metal per cm^2 we may conclude with some degree of certainty that the chemiadsorbed complexes are W-H and W-O respectively. The ortho-para conversion of hydrogen and deuterium and more conclusively the exchange reaction $H_2 + D_2 \rightleftharpoons 2HD$ show that in chemiadsorption at least partial dissociation to form a surface hydride takes place. This method of examination has been extended to investigate the state of chemiadsorbed complex molecules containing hydrogen, such as H_2O, NH_3, and the hydrocarbons. Thus, as shown by Farkas[21], on iron both water and ammonia yield some FeH on chemisorption of the former gases. Likewise, the hydrocarbons undergo disruptive reaction, both at C—H and C—C bonds[22] on chemisorption at metallic surfaces. As already mentioned, the energy changes involved in these chemiadsorptive reactions are still somewhat uncertain. In the case of the chemiadsorption of nitrogen on tungsten the data both from the change in the accommodation coefficient of neon as well as the variation in the contact potential suggest that the primary formation of a surface compound which may be regarded as W_3N proceeds with no or small energy of activation, but that the subsequent formation of the surface compound W-N involves a relatively large energy of activation. We note that a potassium film (lattice constant 5·25 A.) does not react within the temperature range of its stability (600° K.) with molecular hydrogen, although it adsorbs atomic hydrogen with great ease[23]. Carbon (lattice constant 3·4–4·5 A.) reacts with molecular hydrogen and

forms chemiadsorbed hydrocarbon, requiring an energy of activation lying between 13 and 30 kcal. dependent on the surface spacing[24], whilst, as we have seen, tungsten (lattice constant 2·5 A.) reacts with hydrogen even at liquid-air temperatures, although a small activation energy equivalent to the Van der Waals' field of hydrogen appears to be required. From kinetic data we may conclude that the energy of activation of hydrogen on palladium cannot exceed 2·5 kcal./mol., and the energies of activation for the adsorption of hydrogen, ammonia and ethylene on nickel cannot exceed 8 kcal. (Farkas[25]). Garner and Veal[26] find a negligible energy of activation for the chemiadsorption of carbon monoxide on zinc oxide. The bearing of these chemiadsorptive processes on surface catalytic reactions is becoming clearer as a closer study of the kinetics is made.

An important statistical analysis made by Roberts indicates that on adsorption of a diatomic molecule with dissociation, there will be left in the process of covering by the pairs of atoms, some 8 per cent of isolated atoms of metal. Above these isolated atoms existing in the chemiadsorbed primary layer, the field will be stronger than above the monolayer itself. As a result, at suitable pressures and temperatures, a sparsely populated secondary layer will be built up. Such layers have already been observed by Roberts in the case of hydrogen and oxygen on tungsten, covered with chemiadsorbed layers of the respective gases. Finally, at still higher pressures and low temperatures, as was originally shown by Langmuir in the case of oxygen on tungsten, a Van der Waals' secondary molecular layer will be formed on the chemisorbed primary layer of W-O. To give some idea of the energies involved, we may mention that the interaction of molecular oxygen with metallic tungsten to form the primary layer of W-O is accompanied by a thermal change of the order $\Delta H = 128$ kcal. g./mol. The formation of the sparsely populated O_2 layer above the 8 per cent of vacant atoms in the monolayer yields $\Delta H = 45$ kcal./mol., whilst the heat of adsorption of the oxygen on the W-O surface is probably of the order of 10 to 15 kcal./mol.

So far we have assumed that the primary layer formed by disruptive chemisorption of a diatomic gas on a metal is rigid and not mobile. Evidence, however, for activated migration is rapidly accumulating, and in order to work out a complete kinetic picture of the attainment and conditions of equilibrium of a metal with a gas it is necessary to examine all the possible reactions which occur.

In table 1 is given a summary (Barrer) of the published energies of activation for various migratory processes.

The simplest reaction involved is the surface migration of the adatoms which takes place under the influence both of the dispersive forces and of the dipolar repulsive fields between the adatoms. Such surface migration, first demonstrated for metals in the case of thorium in tungsten by Langmuir, and for hydrogen and oxygen on copper and copper oxide respectively in the writer's laboratory, has been examined theoretically by Volmer, Jost, and especially by Lennard-Jones[27]. The adsorbed atom is regarded as lying at the bottom of a potential trough, and on activation it is raised to a migratory level, on deactivation it falls back on to another atom of the substrate. Various assumptions can be made as to the mean free path

Table 1

Helium in silica. Pore diameter (A.)	E (kcal.)	Interdiffusion. Metal E (kcal.)	Gas solid	Ion solid
—	—	—	H in Fe 9·4	K^+ in glass 27·6
3·91	5·70	Pt in Au 39·0	H in Ni 13·8	Na^+ in glass 20·6
4·00	4·19	Cu in Au 27·4	H in Pt 8·0	Na^+ in NaCl 11·8
4·01	3·04	Au in Ag 26·8	H in Cu 14·0	Liquid Liquid
4·20	2·31	Cu in Ni 37·6	O in Cu_2O 10·0	PhOH in MeOH 3·2
—	—	Au in Pb 14·0	O in Ag 21·6	PhOH in C_6H_6 3·1
—	—	Pb in Pb 28·0	N in Mo 45·0	$CaCl_2$ in H_2O 4·4
—	—	Th in W 90·5	N in Fe 23·6	KCl in H_2O 4·0
—	—	C in Fe 33·0	—	D_2O in H_2O 5·3
—	—	Bi in Pb 18·6	—	—

in the migratory level, i.e. the lifetime and horizontal velocity, but no rigid experimental data is available.

Taylor and Langmuir proposed a two-dimensional diffusion equation of the form

$$D' = \tfrac{1}{2} v\lambda$$

for surface migration, where v is the velocity and λ the mean free path of the migrating particle. Jost[38] employed the equation in the form

$$D = \tfrac{1}{2} v\lambda\, e^{-b/T} = KT^{\frac{1}{2}} m^{-\frac{1}{2}} \lambda,$$

on the assumption that there are static (m_s) and mobile (m_m) atoms in equilibrium on the surface. We note that

$$m_m = m_s\, e^{-b/T} f(T),$$

$f(T)$ being a factor dependent on any change in specific heat on transition from a static to a mobile state.

Lennard-Jones, on the other hand, has suggested the form

$$D = \frac{\lambda^2}{2}\tau$$

for the diffusion equation, where τ is the time interval between two successive activations of the adatoms.

If τ be regarded as related to τ_0, the period of vibration of the adatom normal to the surface, by the equation

$$\tau = \tau_0\, e^{-b/T},$$

we obtain

$$D = \frac{\lambda^2}{2\tau_0}\, e^{-b/T}.$$

According to Lennard-Jones τ_0 is of the order of 10^{-12} to 10^{-13} sec. With the experimental values of D, b, and T for sodium on tungsten we obtain $\lambda \sim 18$ A., and for potassium on tungsten $\lambda \sim 40$ A., i.e. the mean free path in flight does not exceed 10 atoms across the surface. It is interesting to note that on this basis gaps of several atomic diameters can be bridged during migration.

If the energy of activation for lateral migration is small it is evident that a surface can be completely filled with atoms by chemisorption of a diatomic gas; for the

8 per cent of voids left on adsorption when the film is immobile will be eliminated by the migratory process. Not only will the saturation maxima differ as the surface film is mobile or immobile, but as shown by Roberts, the kinetics of film formation and the equilibrium will likewise differ. Owing to the normally repulsive fields between the adatoms in a suitably mobile film, the surface tends to become uniform, as collection proceeds, with the result that when the film is half-formed, i.e. $\theta = 0{\cdot}5$, the rise in interaction energy so reduces the probability of two neighbouring atoms becoming simultaneously vacant, and thus permitting of condensation of a gas molecule, that the rate of filling of a surface by condensation becomes appreciably less than for a non-mobile film. A typical table for the ratio of the times of filling of a surface to various extents under the conditions for a mobile film, with interaction energy V, such that $V/kT \doteqdot 9{\cdot}0$ in contrast to an immobile film, is given in table 2.

Table 2

θ	t_2/t_1	θ	t_2/t_1
0·1	1·0	0·6	$2{\cdot}8 \times 10^2$
0·3	1·06	0·7	$9{\cdot}7 \times 10^2$
0·4	1·20	0·8	$1{\cdot}9 \times 10^3$
0·5	2·97	0·9	$1{\cdot}2 \times 10^3$

The filling up of such a mobile film may thus proceed relatively slowly and present several features observed in the phenomenon of activated adsorption.

It is to be noted that if the film be very mobile, i.e. if the time interval elapsing between two migratory excitations be comparable to the lifetime of a molecule adsorbed on a hole, then such a system will fill up more rapidly than either a rigid or relatively mobile film. We might note further that holes can be filled up by atoms diffusing from the interior of the metal. Such complete films present interesting features, such as periods of induction when catalytic reactions are involved.

A distinction between immobile and mobile films possessing interaction energy can be made by determination of the heat of adsorption.

When there is no interaction energy the heat of adsorption on a uniform surface is clearly independent of the surface concentration. If the interaction energy is repulsive, the heat of adsorption will fall as θ increases, and if the film be immobile the fall in ΔH with increase in θ will be linear. Such is the case for the adsorption of hydrogen by tungsten at 280° C., where ΔH falls in a linear manner according to Roberts from $\underset{\theta=0}{\Delta H} = 45{,}000$ cal./g.-mol. to $\underset{\theta=0{\cdot}92}{\Delta H} = 18{,}000$ cal./g.-mol. The various forms of the ΔH, θ curves for mobile films have been examined by Wang.

We may note that if instead of repulsive forces only between the adatoms, attractive as well as dispersive forces are assumed to exist, the surface film may split up into two two-dimensional separate phases of different concentrations. Formal definition of such critical conditions has been made by Frenkel[29] and statistically by Fowler[30]. In the case of the evaporation of sodium from tungsten, Bosworth found that $\Delta H = 32$ kcal. for a very dilute film falling to $\Delta H = 17$ kcal. for a film of minimum work function. This is much less than $\Delta H = 25$ kcal. found for the heat of evaporation of solid sodium; consequently at still higher concentrations ΔH must increase

with increase in the surface concentration. Thus above a certain surface concentration, the surface film is unstable and breaks up into a highly concentrated surface film, and a very sparsely populated one. These conclusions are confirmed by studying the rates of evaporation of concentrated films where it appears the evaporation takes place from the vapour film maintained at saturation by migration from the condensed phase. Examination of such two phase films by the Schottky effect leads one to suppose that the condensed phase is spread uniformly over the apparent surface, and it appears most likely that the condensed phase is present along the edge of the micro-crystals with the vapour phase on the crystal planes.

The electron microscope has recently been used by Johnson and Shockley[31] to obtain an actual picture of the accumulation of adsorbed material along crystal edges during the adsorption of caesium and potassium on tungsten and the activation of thoriated filaments. A clean tungsten wire smoothed by electrolysis or abrasion from all major surface irregularity gave a uniform electron image, showing that, within the resolving power of the microscope, the electron emission was coming uniformly from all points of the wire. The image of a thoriated wire, however, immediately after a flash at 2800° K. consisted of intensely bright spots, many of which reappeared in the same place after subsequent flashes at 2800° K. On heating the wire to an activating temperature of 1850° K., the spots spread out until the whole image was uniformly bright, so that we have by this method, if necessary, a means of taking an actual cinematographic record of the process of surface migration of thorium on tungsten. Johnson and Shockley also observed that a tungsten wire in caesium or potassium vapour, when not carefully polished, gave a patchy image, indicating a work function which varied over the surface. The high-temperature image (surface bare) was always dark where the lower temperature image (surface covered with dilute film of alkali) was bright and vice versa, showing that the alkali is most strongly adsorbed on the points of highest work function, and least adsorbed on the points of lowest work function. It is very satisfactory to obtain an actual photographic record of this sort of behaviour, since so much modern work has pointed indirectly to it.

It is interesting to compare the magnitude of the repulsive forces between the adatoms of potassium on tungsten where the dipole forces predominate with the spreading forces, primarily due to thermal agitation, e.g. formic acid on mercury for similar surface concentrations. The data given in table 3, due to Bosworth, may be cited.

Table 3

Adatoms/cm.² × 10^{-14}	*F* (dyne/cm.)		Activation energy for migration of K on W (kcal. per atom)
	H.COOH on Hg	K on W	
0·12	—	0·96	14·7
0·8	14	26·0	12·1
1·5	22	60	10·3
3·0	38	187	7·2
4·6	52	307	6·6

The dipole moment of the adsorbed potassium varies from $\mu = 23$ debyes at $\theta = 0{\cdot}012$ to $\mu = 2{\cdot}8$ debyes at $\theta = 1{\cdot}0$, showing the large effect of mutual depolarization.

In the case of oxygen adsorbed on the same substrate, the surface-spreading forces due to dipolar repulsion are less than for potassium but are still appreciable, Bosworth's values for the dipole moment of the W-O complex being $\mu = 1{\cdot}0$ debye at $\theta = 0{\cdot}05$ to $\mu = 0{\cdot}8$ debye at $\theta = 0{\cdot}80$, for W-H μ is 0·42 debye. These values can be compared with $\mu = 0{\cdot}2$ debye for palmitic acid molecules at an air-water interface and *ca.* 2·0 debye for a polar group in an organic molecule.

If, as imagined by Lennard-Jones, the surface-energy levels of the adatoms, when mobile, are farther away from the surface than when fixed, it is evident that the mobile adatoms should possess a larger moment than the fixed ones. In the case of sodium on tungsten, we calculate (Bosworth) that at 600° K. a fraction $e^{-1400/600} = 0{\cdot}1$ of the surface adatoms are in the higher level with 0·3 ev. more energy than in the ground state. This fraction increases the contact potential by 0·16 v., representing an increase in dipole moment of 4·7 debyes, or the arm of the electric moment has been increased by 1·0 A., i.e. the mobile level is 0·5 A. farther away from the surface than the bottom of the potential trough.

These migratory processes are not confined to the surface of the metallic substrate, but penetration into the metal nearly always takes place simultaneously. Such penetration has been found to proceed in two ways; first down the intercrystalline boundaries of the micro-crystals, forming the metal or oxide substrate, and into the lattice itself, and the mechanism and kinetics of these processes bear not only on the problem of adsorption, but on persorption and diffusion processes in metals.

As is well known, especially from the work of Sieverts[32], many diatomic molecules dissolve in metals in the atomic form, leading to the well-known Richardson $p^{\frac{1}{2}}$ law for the rate of solution. The dissolved gas may be in the form of free protons, e.g. in Cu, Ni, Co, Fe or in the form of hydrides, e.g. Zr,Ti. Theoretical formulae for the solubilities have been elaborated by Fowler and Smithells[33]. Investigation on the rates of solution have been made by Dunn, Ward[34] and Melville in the writer's laboratory[35], and by Smithells and Ransley[36]. In all cases it is found that the diffusive process inside the metal, whether extra- or intra-lattice, is of the activated type, and in the case of the diffusion of oxygen through cuprous oxide, these two diffusive processes are associated with very different energies of activation.

In the case of the penetration of oxygen through cuprous oxide to an underlying copper substrate, Pilling and Bedworth showed that the rate of oxidation became independent of the pressure above a certain critical pressure. This critical pressure, however, varies with the state of crystallization of the cuprous oxide surface. When extremely finely divided, the critical pressure is high, but when the surface consists of macro-crystals, the critical pressure is extremely low (Wilkins[37]). We must conclude that the primary act in the penetration process is the adsorption of the oxygen on the cuprous oxide surface, and the lateral migration of the adsorbed gas to the intercrystalline boundaries. This lateral migration is of the activated type.

If we turn to the passage of hydrogen into metals such as nickel, palladium and copper, here again atomic activated diffusion in the interior represents the movement in the bulk phase, but the mechanism of transition from the gas to the bulk metal phase has not been completely elucidated. We have seen that in chemisorption on the surface, dissociation into atoms takes place, resulting in the formation of a surface hydride. This process involves either a negligibly small energy of activation, as in the case for hydrogen on tungsten, or a relatively large one, as in the case for hydrogen on carbon[38]. The transfer of such an adsorbed atom into the interior involves, according to Melville and Rideal, another energy of activation, and finally the migratory passage in the interior involves still a third energy of activation. If the surface phase were the result of only weak adsorption and penetration therefrom involved a high energy, then the surface atom composition θ would be related to the bulk gas molecular concentration by the equation $\theta^2 = kp$ and Richardson's equation would be valid, except at very low pressures, where the rate of uptake of hydrogen by the metal would be governed by the rate of collision of the gas with the surface. As we have noted, however, in the case of tungsten at least, chemisorption takes place with extreme ease and involves a high heat of adsorption, and it is only at extremely low pressures that the equation $\theta^2 = kp$ can be valid, yet as shown by Smithells and Ransley, the Richardson equation holds true in the case of nickel to over 100 atmospheres. It is possible that the 8 per cent of vacant spaces originally present in the chemisorbed layer play an important part in the penetration process. We may regard the mechanism of penetration in two different ways. Either penetration takes place both from the chemisorbed layer, but much more readily from the molecular layer above the holes in the chemisorbed layer, or only molecules with sufficient energy of activation to penetrate the surface layer actually do so. Atomic adsorption may take place from a molecule retained above a hole provided that one atom of the molecule simultaneously undergoes penetration. Such a process must be reversible, i.e. atoms diffusing outward from the interior of the metal must combine with a surface-chemisorbed atom to form both a hole and a molecule above the hole. On this view holes are being continuously formed and destroyed. The dynamics of such a reversible process have been elaborated by Wang[39].

If we consider the case of capillary flow, we can distinguish three cases: turbulent and stream-line flow at high pressures investigated by Poiseuille and Reynolds; low-pressure flow with reflexion, condensation and evaporation at the walls and free flight in the gas phase (λ, d comparable), i.e. the so-called Knudsen flow, investigated by Clausing; and finally, flow in constricted capillaries or in lattices, migration proceeding by passage over a periodic series of potential barriers. In this latter case, as we have seen, the migrating molecule is never free, but is merely elevated into its transitional potential level by activation. It would appear that the energy of activation for migration is in general some one-third of the energy required for evaporation on an external surface.

The mechanism of the process of transition from a chemisorbed monolayer to a thick layer of condensed solid or liquid has not yet been investigated in a satis-

factory manner. As in films on liquid surfaces, the transition from a two-dimensional vapour to a two-dimensional condensed phase takes place on solid substrates, although the experimental data appear to be only satisfactory in the case of the alkali metals on tungsten. We have noted that secondary layers can be built up on chemi-adsorbed monolayers of the gases, such as oxygen on WO, hydrogen on WH and the like, but for vapours the progressive stages of condensation are not so clearly defined. It appears, for example, that when a bimolecular layer on iodine is built up on potassium iodide, the new surface has all the characteristics of solid iodine, i.e. a mobile surface phase. The same appears to be true of other vapours, below their critical point. When attempts are made to increase the thickness of the layer above that corresponding to two molecules in many cases (i.e. in liquids), the layers break up with small drops showing the so-called dew-point phenomenon. For solids at not so low temperatures, micro-crystals are formed by lateral migration. In this case, at least, the instability of the thick layer is considerably enhanced by the presence of inclusions.

At a very early stage in the consideration of adsorptive and catalytic actions at solid surfaces, hypotheses have been made involving the non-uniformity of the solid surface in respect to its adsorptive and catalytic powers. According to Taylor, isolated atoms on the surface represented points of strong attachment; the writer expressed the view that the edges of the micro-crystals were the cause of the apparent differentiation in properties, whilst Balandin was the first to suggest that certain unique atomic spacings present in the metal surface were the active patches. As we have noted more recently, both the electron microscope and the Schottky effect reveal the lack of uniformity of the field at high temperatures on a tungsten surface. Conclusions previously based on the changes in the heats of adsorption have now to be accepted with caution until the magnitude of the dispersive forces has been properly evaluated. Maxted[40] in a lengthy series of papers on the effects of poisons on catalytic activity, concludes that catalytic surfaces are uniform in respect to their activity. It might be remarked that if, as is the case in Maxted's experiments, a strong poison, i.e. a molecule with a very long lifetime in the adsorbed state, is employed, its chances of migration from a low to a high field on the surface is small, and in consequence a uniform deposition and thus distribution of poison on the surface results. Different results might be obtained if an inhibitor less strongly adsorbed were employed, for here, in spite of uniform deposition, enhanced evaporation and migration from areas of weak fields would ensue and active patches might be revealed.

It is interesting to note in this connexion that chloroform does not exert an identical inhibitory effect on the two methods of decomposition of alcohol, at an alumina surface, suggesting that two different types of adsorbing surface are involved.

Increasing use has been made of the hydrogen isotope deuterium for investigations on the mechanisms of reaction at solid surfaces. If the rates of reaction of hydrogen and of deuterium were determined only by a collision factor involving the masses, we would anticipate that the relative rates of reaction under comparable

conditions would be $\sqrt{2}$ or 1·414 : 1. In addition to the mass difference, however, the zero-point energies of the two molecular species are different, being 6·1 kcal. for H_2 and 4·4 kcal. for D_2[41]. Thus we might anticipate that in those reactions where the velocities are governed by the rate and energy of activation of molecular hydrogen, the relative velocities should be capable of being interpreted in terms of this difference in activation energy. Again, the sequence of reactions involved may proceed through a series of steps, e.g. $H_2 \rightarrow MH \rightarrow M + HX$, and the difference between the energies of activation of the reactions involving the isotopes may involve only the difference between the zero-point energies of the surface hydrides. These should not differ very considerably from the values obtained from band spectroscopic data on the volatile metallic hydrides, and Eyring and Sherman[42] have estimated this difference between NiH and NiD as 0·7 kcal., or only one-half of that obtaining for the molecules. It might be noted, however, as first indicated by Polanyi[43], that if the process of the reaction be imagined as one involving the excitation of the reactants to an activated complex, the difference in zero-point energies between the two activated complexes may not be identical with the difference existing between the zero-point energies of the reactants, and as a result the net difference in energies of activation between the two reactions will be less than that anticipated by considering the zero-point energies of the reactants alone.

In addition to such energetic considerations, deuterium has found application in investigating the mechanism of catalytic interactions. Many cases of homogeneous exchange reactions are now known in which hydrogen in organic or inorganic molecules is found to be replaceable by deuterium when a deuterium-containing compound is brought into contact with the latter. For exchange reactions in non-ionizable systems the reaction velocity is accelerated by a catalyst. The first heterogeneous reaction of this character was established by Farkas and Rideal[44], who showed that in the presence of nickel the interaction between ethylene and deuterium involved not only the addition reaction to form dideuteroethane, but also an exchange reaction of the type $C_2H_4 + D_2 \rightleftharpoons C_2H_3D + HD$. Exchange reactions in other unsaturated hydrocarbons have been examined by Polanyi and his co-workers[45], and extended to the saturated hydrocarbons by Taylor[46] (cf. Morikawa, Benedict and Taylor). Comparisons of rates in the liquid and vapourphase and the examination of other exchange reactions have been made by Farkas[47] and by Farkas and Farkas[48].

In the case of the unsaturated hydrocarbons, it has been shown that the exchange and addition reactions proceed by different mechanisms; the former proceeds certainly in atomic states, whilst the latter appears to consist in the saturation of a bond with a molecule or two atoms. The evidence that hydrogenation of olefines at catalytic surfaces is a local, possibly molecular, and not an atomic process, rests on the observation that 1·2 saturation and never 1·4 reaction occurs on hydrogenation. The evidence is not really quite conclusive, in view of the fact that dissociation as well as addition takes place, unless the reactivity of the groups after addition of only one atom is regarded as unaltered. In respect to the exchange reaction we may adopt the view that this results from a prior addition of an atom of

deuterium to one carbon atom in the double bond, with a subsequent loss of one of hydrogen from the saturated group, e.g. $=CH_2 \rightarrow -CH_2D \rightarrow =CHD$, or we may imagine the reverse sequence of affairs, i.e. loss followed by addition

$$=CH_2 \rightarrow =CH- \rightarrow =CHD.$$

Deductions as to the sequence of events from either the energetics or pressure relationships of the reaction velocity are at present somewhat conflicting, as somewhat arbitrary hypotheses have to be introduced as to which is the rate-determining step.

In the case of exchange reactions in the saturated hydrocarbons it is most probable that the dissociation stage precedes addition, and according to Taylor's measurements this activated dissociation is the rate-controlling step, the energies involved being of the order of 19 kcal. for the C—H and some 43 kcal. for the C—C link for nickel. The apparent activation energies on platinum for the hydrogenation and exchange reactions of benzene are lower and only some 7 to 9 kcals. In a surface complex reaction of the type $MXH \rightleftharpoons MX + MH$ the difference in energies of activation for the forward and back reaction will clearly be dependent on the magnitude of the heat of formation of the surface compound, from the surface reactants.

It might thus be anticipated from the above data that the heat of formation in a purely surface reaction of a C—H link is of the order of $19-7$, e.g. 12 kcal. This value can be compared with the reduction of a surface amide by hydrogen on iron of 15 kcal.[45], whilst on the same metal the apparent energies of activation for the ortho-para hydrogen, the ortho-para deuterium and the hydrogen-deuterium reactions are 8·1, 8·4 and 9·0 kcal. respectively. It must be observed, however, that several interpretations of these values in terms of molecular mechanisms may be made, and definite conclusions must await further experimental work.

We may include the figures for a reduction of a surface oxide or hydroxide on platinum of 13·5 kcal. and 16·4 kcal. on nickel[40].

REFERENCES

(1) Benton, Taylor. *Trans. Faraday Soc.* 28 (1932).
(2) Rideal. *Trans. Faraday Soc.* 28 (1932).
(3) Roberts. *Proc. roy. Soc.* A, **129**, 146 (1930).
(4) Langmuir. *Acta phys. Chem. U.R.S.S.* **1**, 34 (1934).
(5) Becker. *Trans. Amer. electrochem. Soc.* **55**, 153 (1929).
(6) de Boer. *Physica*, **2**, 114 (1935).
(7) Bosworth. *Proc. roy. Soc.* A, **112**, 154 (1936).
(8) Bosworth and Rideal. *Proc. roy. Soc.* A, **162**, 1 (1937).
(9) Jackson and Mott. *Proc. roy. Soc.* A, **137**, 703 (1932).
(10) Jackson and Howarth. *Proc. roy. Soc.* A, **142**, 447 (1933); A, **152**, 515 (1935).
(11) Landon. *Phys. Z. Sowjet.* **8**, 489 (1935).
(12) Volmer. *Z. phys. Chem.* **115**, 239.
(13) Fowler. *Proc. Cam. phil. Soc.* **31**, 210 (1935).
(14) Lennard-Jones, Strachan and Devonshire. *Proc. roy. Soc.* A, **150**, 442, 456 (1935); A, **156**, 6, 29 (1936).
(15) Lennard-Jones and Devonshire. *Proc. roy. Soc.* A, **158**, 242 (1937).
(16) Fowler. See (13).

(17) PEIERLS. *Proc. Camb. phil. Soc.* **32**, 471 (1936).
(18) ROBERTS. *Proc. roy. Soc.* A, **152**, 445 (1935).
(19) TOPPING. *Proc. roy. Soc.* A, **111**, 62 (1932).
(20) MELVILLE. *Sci. Prog.* **121**, 107 (1936).
(21) FARKAS. *Trans. Faraday Soc.* **32**, 425 (1931).
(22) TAYLOR. *J. Amer. chem. Soc.* **55**, 1445 (1936).
(23) BOSWORTH. *Trans. Faraday Soc.* **32**, 1370 (1936).
(24) BARRER and RIDEAL. *Proc. roy. Soc.* A, **149**, 231 (1935).
(25) FARKAS. *Trans. Faraday Soc.* **32**, 475 (1936).
(26) GARNER and VEAL. *J. chem. Soc.* p. 1495 (1935).
(27) LENNARD-JONES. *Trans. Faraday Soc.* **28** (1932).
(28) JOST. *J. chem. Phys.* **1**, 466 (1933).
(29) FRENKEL. *Z. Phys.* **26**, 117 (1934).
(30) FOWLER. *Proc. Camb. phil. Soc.* **31**, 260 (1935).
(31) JOHNSON and SHOCKLEY. *Phys. Rev.* (1936).
(32) SIEVERTS. *Z. phys. Chem.* **60**, 129 (1902).
(33) FOWLER and SMITHELLS. *Proc. roy. Soc.* A, **160**, 37 (1937).
(34) DUNN and WARD. *Proc. roy. Soc.* A, **133**, 506 (1931).
(35) MELVILLE and RIDEAL. *Proc. roy. Soc.* A, **133**, 89 (1935).
(36) SMITHELLS and RANSLEY. *Proc. roy. Soc.* A, **150**, 172, 152, 706 (1935); A, **155**, 195 (1936).
(37) WILKINS. *Proc. roy. Soc.* A, **128**, 407 (1930).
(38) BARRER. *Proc. roy. Soc.* A, **149**, 253 (1935).
(39) WANG. *Proc. Camb. phil. Soc.* **32**, 657 (1936).
(40) MAXTED. *J. chem. Soc.* **26** (1934 *et seq.*).
(41) UREY and RITTERBERG. *J. chem. Phys.* **1**, 137 (1933).
(42) EYRING and SHERMAN. *J. chem. Phys.* **1**, 348 (1933).
(43) POLANYI. *Proc. roy. Soc.* A, **144**, 16 (1934).
(44) FARKAS and RIDEAL. *Proc. roy. Soc.* A, **146**, 630 (1934).
(45) POLANYI. *Trans. Faraday Soc.* **30**, 603, 1164 (1934).
(46) TAYLOR. *J. Amer. chem. Soc.* **55**, 1445 (1936); **58**, 1795 (1936).
(47) FARKAS. *Trans. Faraday Soc.* **178**, 416 (1936).
(48) FARKAS and FARKAS. *Trans. Faraday Soc.* **30**, 1071 (1936).
(49) MELVILLE. *J. chem. Soc.* **26**, 797 (1934).

SURFACE TENSION

By R. C. BROWN, Ph.D.

§ 1. INTRODUCTION

It is not intended in this article to make a comprehensive survey of our present knowledge of surface phenomena. The study of adsorption and thin films is, rightly or wrongly, looked upon by the physicist as lying almost completely within the domain of the physical chemist—one rarely meets a chemical physicist. This state of affairs, together with the fact that articles on surface phenomena are to be found in the *Annual Reports* of the Chemical Society, has inclined the writer to concentrate on the purely physical aspects of surface tension. Inevitably some mention of adsorption is made in the section dealing with solutions, but if any justification is needed for its inclusion it must be made on the grounds of general interest.

§ 2. THEORIES OF SURFACE TENSION

The last few years have seen a number of attempts to modify classical theories of surface tension in the light of modern ideas concerning the structure of liquids.

R. H. Fowler[1] has put forward a noteworthy "tentative statistical theory" of Macleod's equation and the parachor.

It will be remembered that Macleod's equation is simply

$$\sigma = \text{const.}\ (\rho_{\text{liq}} - \rho_{\text{vap}})^4,$$

where σ stands for surface tension and ρ for density. The parachor, P, is defined by

$$P = \frac{M\sigma^{\frac{1}{4}}}{\rho_{\text{liq}} - \rho_{\text{vap}}},$$

where M is the chemical molecular weight of the liquid.

Fowler denotes by ΔE_σ the "excess in the average configurational energy of a system in its equilibrium state with an extra two units of area of interface over the value for the system in a standard equilibrium state". This excess can be looked upon as a contribution to the average total energy and can be related to the increase of free surface energy, 2σ, by the Gibbs-Helmholtz equation

$$-\frac{\Delta E_\sigma}{T^2} = 2\frac{\partial}{\partial T}\left(\frac{\sigma}{T}\right) \qquad \text{.....(1),}$$

in which T is absolute temperature.

Fowler then points out that σ vanishes when T reaches the critical temperature, T_c, so that

$$\frac{\sigma}{T} = \frac{1}{2}\int_T^{T_c} \frac{\Delta E_\sigma}{T'^2} dT' \qquad \text{.....(2).}$$

Rayleigh's classical theory of capillarity is then modified by the introduction of the neighbour-distribution function $g(r)$, which expresses the average number of molecules in a particular region when a chosen molecule is at a certain point, distance being represented by r. An expression for the average total interaction energy between a semi-infinite volume and a semi-infinite column of liquid is thus obtained and used in order to calculate ΔE_σ for a liquid-vapour interface. This is done by imagining that a column of liquid is broken to form two units of new surface and that the same is done for a column of vapour. The two pairs of liquid-vapour surfaces are then placed together and the net increase of energy resulting from the whole process is evaluated. It is necessary to make assumptions (valid near the critical temperature) with regard to the functions like $g(r)$ and their dependence on temperature. Further it is assumed that the liquid and vapour densities near the critical point obey a relation such as

$$(\rho_{\text{liq}}-\rho_{\text{vap}})^2 = A(T_c - T),$$

where A is a constant.

Then it is found that

$$\Delta E_\sigma = B(T_c - T) \qquad \text{.....(3)},$$

where B is effectively constant near $T = T_c$. Substitution in equation (2) and further simplification yields Macleod's equation or that for the parachor, thus

$$P^4 = \frac{B}{4A^2T_c}N^4 \qquad \text{.....(4)},$$

where N is Avogadro's number.

In order to make a numerical test of equation (4) Fowler makes further simplifications to the factor B and the quantities which it finally involves are the interaction energy between a pair of molecules, the molecular diameter, d (such that when $r < d$, $g(r) \equiv 0$), and lastly the neighbour distribution function.

For argon the agreement between the experimental and calculated values of P is quite good, the values being 54 and 62 respectively. For mercury the corresponding values are 68·5 and 53, which is considered to be less satisfactory since Fowler expects his theory to give a somewhat high value on account of the nature of the assumptions made.

The paper concludes with a few suggestions concerning the rather remarkable fact that although throughout the analysis the temperature of the liquid is supposed never to be far removed from T_c yet in actual fact Macleod's equation holds good over a wide range of temperature.

At the conclusion of a paper on the theory of the solid state Weng Wen-Po[2] extends his theory of metals to derive a relation between surface tension at the melting point, σ_m, the infra-red characteristic frequency, ν, of the substance in the solid state and W its atomic weight.

He finds

$$\sigma_m = 5{\cdot}0 \times 10^{-25}\nu^2 W,$$

and when this is slightly modified he arrives at a relation which gives fairly good agreement with experiment. He further suggests that for all monatomic liquids the

expression $\rho^{\frac{2}{3}} T_m W^{-\frac{2}{3}} \sigma_m$, where T_m is the melting point, should be a constant. This again is fairly in agreement with observation, the constant being approximately unity.

G. Bakker[3] has continued his theory of the labile portion of the isotherm and its relation to the capillary layer.

§ 3. THE SURFACE TENSION OF SOLUTIONS

A large amount of work is being done on this subject, the greater part of it being of an experimental nature. The number of different mixtures and solutions which can be made is clearly legion, and it will be a long time before all their physical properties are determined.

As far as aqueous solutions are concerned it will be remembered that substances may be divided roughly into two classes: (1) capillary-inactive substances, i.e. those which raise the surface tension of water and therefore have, according to the theory of Willard Gibbs, a lower concentration in the surface than in the bulk (negative adsorption), and (2) capillary-active substances which lower the surface tension of water and are therefore positively adsorbed. Inorganic salts form the largest group in class (1), and in class (2) are to be found organic acids, alcohols, esters and so on.

Theoretical. Theories of the behaviour of electrolytes were published by Wagner[4] in 1924 and by Onsager and Samaras[5] in 1934. They gave the following equation for the surface tension, σ_c, of a dilute solution of concentration c in terms of that of pure water, σ_0, and the dielectric constant of water, D:

$$\frac{\sigma_c}{\sigma_0} = 1 + \frac{79{\cdot}517}{D\sigma_0} \log \frac{1{\cdot}143 \times 10^{-3} (DT)^3}{c} \qquad \text{.....(5).}$$

K. Ariyama[6] criticizes the above theories chiefly on account of their mathematical difficulty and suggests a comparatively simple treatment.

According to the Debye-Hückel theory any ion is surrounded by an ion atmosphere which, on the average, has a charge opposite in sign to its own, and the potential energy of such a single ion is given by $-E^2 m/D$, where E is the ionic charge, D is the dielectric constant of the solution and $m^2 = 8\pi C_\infty E^2/DkT$, C_∞ being the average bulk concentration of the solution and k the Boltzmann constant.

Thus it is suggested by Ariyama that the ion atmosphere may be replaced by a sphere of radius $a = 1/m$ upon whose surface the charge $-E$ is uniformly distributed. This would give the same potential $(-E/aD)$ at the centre as does the ion atmosphere, and it may be treated as the sphere of influence of the classical theories of surface tension.

An ion situated at a depth less than a below the surface has an incomplete sphere of influence and consequently its potential energy is different from that of an ion in the bulk of the solution. Using this idea together with the Boltzmann equation for the relation between the concentrations at any two points and the energies of the particles at those points, Ariyama is able to obtain an approximate expression for the excess concentration of ions in the surface as follows:

$$\alpha = -\frac{C_\infty}{kT} \frac{E^2}{4D} \qquad \text{.....(6).}$$

Comparing this with the Gibbs equation

$$\alpha = -\frac{C_\infty}{RT}\left(\frac{\partial\sigma}{\partial c}\right)_T \qquad \text{.....(7),}$$

we see that

$$\left(\frac{\partial\sigma}{\partial c}\right)_T = \frac{E^2N}{4D} \qquad \text{.....(8).}$$

This equation suggests that, at any rate for dilute solutions, the surface tension should increase with concentration at the same rate for all electrolytes, and this is in fact approximately the case. For higher concentrations the rate of increase of surface tension actually diminishes. In a second paper [7] a slightly different approach is made to the same problem.

Ariyama [8] has pointed out also that inorganic acids do not behave in the same way as salts inasmuch as they lower the surface tension of water, i.e. they must be positively adsorbed at the surface. This suggests immediately that the H^+ ions which are present in the acid solutions are the cause of the different behaviour, and Ariyama puts forward the idea that a water surface consists of molecules oriented in such a way that their hydrogen-attracting ends are directed inwards. These supply the force which causes the migration of the H^+ ions to the surface with the resulting lowering of surface tension.

In connection with this Ariyama has made a simple calculation which shows the effect of the H^+ ions. The surface tension of a half-normal solution of sodium chloride exceeds that of pure water by 0·83 dyne/cm. This is assumed to be due to the combined equal effects of the Na^+ and Cl^- ions since both have numerically the same charge and approximately equal volumes. Thus the effect of each is 0·41 dyne/cm. A half-normal solution of hydrochloric acid gives a decrease of 0·16 dyne/cm. and, if the contribution of the Cl^- is the same in this case, that of the H^+ must be $-0{\cdot}16-0{\cdot}41=\cdot-057$ dyne/cm. Figures closely agreeing with this one are obtained in a similar way for the effect of the H^+ ions in other acids.

Experimental: (1) *Capillary-inactive solutes.* Some important new facts have been brought to light with regard to the surface tensions of aqueous solutions of electrolytes by Jones and Ray [9]. These authors were concerned with dilute solutions and set out to measure the ratio of the surface tension of the solution, σ_c, to that of pure water, σ_0.

Their apparatus was essentially a U-tube having one limb of capillary dimensions and the other comparatively wide. A mark was made on the capillary and a certain volume of water was put into the apparatus so that the meniscus stood at the mark. The same was done with the solution, and, on account of the (slightly) differing surface tensions, the levels in the wide tube were different in the two cases. The difference of level could be determined by weighing the two quantities of liquid and measuring their densities. Thus the principal term in the expression for σ_c/σ_0 involves densities only and the accuracy is considerable; in fact results are given to five places of decimals.

The results show that for dilute solutions (below 0·006 N) a *lowering* of surface tension occurs, a minimum value being observed at about 0·001 N. This applies to

all the electrolytes used, namely potassium chloride, potassium sulphate and cæsium nitrate. The minimum surface tension is about 0·02 per cent. below that of water; but it occurs at slightly different concentrations in the three cases. Jones and Ray are convinced of the reality of this effect since the lowering observed is about nine times as large as their estimated probable error. Sucrose solutions showed no such minimum—simply a steady rise of surface tension with concentration, and the authors conclude that the positive adsorption which must accompany the initial downward slope of the electrolyte surface tension-concentration curve is dependent upon the electric charges of ions. The effect is not forecast by any existing mathematical theory, e.g. equation (5), because, in the opinion of Jones and Ray, the originators of these theories have concerned themselves only with inter-ionic forces which must at low concentrations be very small compared with the forces which the water molecules exert on each other and on isolated ions.

(2) *Capillary-active solutes.* There is a good deal to report in connexion with solutions of capillary-active substances, and here again we find some new ideas put forward to explain hitherto unnoticed paradoxes. First of all a short account of a new technique of experimentation may not be out of place.

It has long been known that the surface tensions of very dilute aqueous solutions of certain long-chain compounds, soaps for instance, decrease with time, beginning at the value for pure water and reaching an equilibrium value in a time which increases with dilution and may run into several hours or even days. The determination of the maximum pull on a horizontal ring which is slowly raised from the liquid surface has been used extensively for the investigation of this phenomenon. In choosing a method it must be remembered, however, that the contact angle as well as the surface tension may vary with time and this possibility, failing evidence to the contrary, should rule out any method of measuring surface tension which depends on a knowledge of the contact angle. Bearing in mind also that a truly static method is required we are driven to the conclusion that, of the classical methods, the sessile bubble or drop seems to be the most suitable.

Recently, however, the use of the Langmuir-Adam trough has been suggested for the measurement of the surface tension of adsorbed films of soluble substances. This apparatus consists of a shallow rectangular tank filled with liquid in which a transverse barrier floats so as to divide the surface into two parts. If an insoluble film is placed on the liquid surface on one side of the barrier the latter experiences a force equal to the so-called surface pressure of the film multiplied by the length of the barrier. This force is measured by means of a torsion balance. The pressure of a surface is thus equal to the difference between the surface tension of the pure liquid on the one side of the barrier and that of the surface carrying the film on the other. The area of the film may be varied by moving a second barrier which extends right across the trough.

McBain and his co-workers[10, 11] have described their use of a trough apparatus in which the floating barrier divides the actual trough into two sections so that the liquids on either side of it cannot mix, the barrier being attached to the sides and bottom of the trough by means of a thin film of rubber. Thus the solution may be

placed on one side and pure water on the other. The force on the barrier then yields the surface tension of the solution if that of pure water is known, and the time changes may be studied without disturbance of the surface.

Doss[12] also describes a similar technique, but in his case the floating barrier of the original thin-film apparatus is retained and he compares the surface pressure of an ageing surface with that of a surface of known age (or a completely aged one) on the other side of the barrier.

Both sets of workers are concerned to explain the extremely slow rate of adsorption which occurs at low concentrations. Doss dismisses as untenable the earlier idea that the rate of diffusion is slow simply on account of the high molecular weight of the solute, and he suggests that there may exist below the surface a potential barrier the nature of which is somewhat hypothetical. Diffusion into the surface is thus retarded. He finds support for this idea in his observation that if a surface is reduced in area its pressure rises owing to overcrowding, thus indicating a reluctance on the part of the molecules to return to the bulk.

McBain, Ford and Wilson[11] discuss another aspect of the properties of a soap-solution surface. If we always deal with solutions which have attained equilibrium and construct a surface tension-concentration curve it shows a rapid fall from the pure-water value, passes through a minimum and then rises slowly. According to Gibbs's equation, therefore, the surface concentration is the same as the bulk concentration when the surface tension is a minimum, and for higher bulk concentrations the solute molecules are *negatively* adsorbed ($\partial\sigma/\partial c$ positive). In this region, therefore, we have a surface containing more water than soap while its surface tension is very much less than that of water. The authors therefore make the suggestion that immediately after the formation of the surface, the solute molecules already in it become oriented so as to produce the almost instantaneous lowering of surface tension, while those molecules very near the surface move up and leave a layer of water molecules immediately beneath them. The Gibbs adsorption of water then takes place towards this layer of almost pure water, and, moreover, takes many hours for completion.

While discussing the mechanism of adsorption it may not be out of place to refer to the bold but extremely direct method of testing Gibbs's equation which has been devised by McBain and Swain[13]. The solution is contained in a tank and at a given time after the sweeping of its surface a microtome blade travelling at 35 ft./sec. shaves off the top of the liquid to a depth of 0·05 or 0·1 mm. The concentration of the solution which is thus removed is determined by an interferometer method. The tests of the Gibbs equation are found to be much more satisfactory than those based upon the passage of air bubbles through the solution, since in these latter the time during which the adsorption process takes place is necessarily small. Perfect agreement is found in the cases of hydrocinnamic acid, phenol and sodium chloride solutions. Incidentally it may be noted that it is from such experiments[14] that we learn that in the case of the hydrocinnamic acid the composition of a very young surface is the same as that of the bulk (net adsorption = 0) even though the surface tension is perhaps 10 dynes/cm. below that of water. This

fact supports McBain's suggestion of "pre-Gibbs adsorption by surface rearrangement" referred to in the previous paragraph.

To explain the upward trend of the surface tension-concentration curve McBain appeals to dissociation of the surface molecules of solute. This, it is suggested, results in the formation of an electrical double layer which decreases surface tension, and the degree of dissociation decreases with rising concentration.

Homologous series. The properties of solutions of homologous series receive attention in a lengthy paper by Merckel[15]. In dealing with surface tension the author covers ground similar to that opened up by Langmuir[16] in 1917. For instance, the change of the potential energy when a mol of solute moves from the interior of the solution to its surface is found to increase by a constant amount for each addition of CH_2 to the molecule. Again if the initial downward slope $[(\partial\sigma/\partial c)_{c=0}]$ is read off from the surface tension-concentration curve for each member of a series and is denoted by Δ then it can be shown simply by plotting log Δ against n, the number of carbon atoms in the molecule, that

$$\log \Delta = an + b,$$

where a and b are constants for any particular series. By using reasoning similar to that of Langmuir it is found that for any pair of consecutive members of a series

$$\frac{\Delta_{n+1}}{\Delta_n} = e^a.$$

The constancy of the ratio is expressed by Traube's rule, and its value is about 3·4 for the early members of the fatty acid series.

Szyszkowki's equation which represents the surface tension-concentration curve for any of these acids may be mentioned here. It is

$$\frac{\sigma_0 - \sigma_c}{\sigma_0} = B \log\left(\frac{c}{A} + 1\right).$$

B is found to have the same value for all the acids but A is a constant only for each acid, and $A_n/A_{n+1} = 3{\cdot}4$—clearly another way of expressing Traube's rule.

The foregoing observations and empirical relations have led to the conclusion that the acid molecules are, at any rate in dilute solutions, oriented parallel to the surface.

Merckel pursues an analysis of concentrated solutions on similar lines. He assigns a "capillary active number", K, to each member of an homologous series. Thus for the first member ($n = 1$) he puts $K = 0$, and when $n = 2, 3, 4, 5$, $K = 1, 3, 6, 10$ and so on. Then for any series, whether it be fatty acid, alcohol or ester, if the surface tensions at definite chosen concentrations are measured for each member of the series and plotted against the appropriate capillary active numbers the result is a straight line having a negative slope. If a higher concentration is chosen the slope becomes steeper. Surface tension is therefore a linear function of K and, bearing in mind the relation between K and n, we find

$$\sigma = \alpha\frac{n^2}{2} - \alpha\frac{n}{2} + \beta,$$

from which the surface tension at a particularly (high) concentration can be calculated for any member of an homologous series.

Merckel analyses some results of Lottermoser and Winter[17] for the interfacial tensions of solutions of the sodium salts of the fatty acids against paraffin oil. He shows that for these $\log \Delta$ is a linear function not of n but of K; thus

$$\log \Delta = aK + b,$$

and this leads to

$$\frac{\Delta_{n+1}}{\Delta_n} = e^{pn},$$

so that Traube's rule is not obeyed and it also follows that the effect of an additional CH_2 upon the change of potential energy when a mol passes from the surface to the interior is not constant but is proportional to n.

The validity of Traube's rule for a liquid-liquid interface has been investigated by Sonja Boas-Traube and M. Volmer[18] using aqueous solutions of the first six fatty acids against benzene. The rule is found to hold good except perhaps for the earliest members of the series. In addition when the constant, B, of Szyszkowski's equation is calculated for each acid, higher values than those relating to the corresponding solution-air interface are found. This means a greater relative lowering of surface tension and consequently a closer packing of acid molecules at the solution-benzene interface as compared with the solution-air interface.

§ 4. THE SURFACE ENERGY OF SOLIDS—WETTING

If the angle of contact, θ, between a liquid and a solid in the presence of air (or another liquid) is measured in the liquid phase it is fairly obvious that the effectiveness of the liquid as a wetting agent on the particular solid will be greater the smaller θ becomes. In practice, if σ denotes surface tension or free surface energy and the subscripts 1 and 2 refer to solid and liquid respectively, then the adhesion tension, A, is used as a measure of the wetting power and is defined as follows:

$$A = \sigma_1 - \sigma_{12} = \sigma_2 \cos \theta.$$

Clearly A varies between σ_2 and $-\sigma_2$ as θ increases from 0 to 180°. Thus the estimation of the wetting properties of a liquid for a given solid becomes chiefly a matter of contact angle measurement.

The tilting-plate method which has the merit of being direct and simple is used by Wenzel[19], who is particularly interested in the proofing of fabric and affixes strips of various materials on to a microscope slide which is capable of being tilted.

Wenzel emphasizes a fact which seems to be of considerable importance, namely that the degree of wetting of a particular solid by, say, water depends not only upon the composition of the solid but also upon the texture or polish of its surface. Clearly in the case of a roughened or pitted surface a geometrical square cm. may present a considerably greater effective area to the liquid. He therefore defines a number, r, which he calls the "roughness factor" as (actual surface) ÷ (geometrical

surface). The roughness factor, therefore, is unity for a perfectly polished surface but exceeds this value for all others. We may now write

$$rA = r(\sigma_1 - \sigma_{12}) = \sigma_2 \cos\theta.$$

The numerical value of rA exceeds that of A, and consequently if a liquid wets a polished surface ($\theta < 90°$) it will wet a roughened surface to a greater extent, but if the wetting of the polished surface takes place with difficulty ($\theta > 90°$) then the resistance to wetting of a rough surface will be even greater for in this case A is negative.

Wenzel has determined the value of rA for various materials, and points out that the effective proofing of a fabric is aided by its fibrous nature, which gives it a rough surface.

Bartell and Zuidema[20] have used the method of small sessile drops devised by Mack for the measurement of contact angles (see last year's *Progress Reports*). They find, among other things, that soft solids such as talc are wetted with greater difficulty than are brittle substances. This may be explained by their smaller cohesion, which means a lower solid surface tension.

Bartell and his co-workers have published three consecutive papers[21,22,23] on the effect of heat treatment on the free surface energy of solids. The experiments consist essentially of the measurement of contact angles. A vertical rod of solid is dipped into a vessel containing water or an organic liquid which is lying above or below water, and it is claimed that a magnified photograph of the shape of the meniscus near the rod yields reproducible values of the contact angle for either of the systems solid-water-air or solid-water-organic liquid. In order to study the effect of heat treatment on the solid it was necessary to be able to reproduce a standard surface, and it was found that a particular series of operations involving scraping, polishing, washing and drying produced surfaces of gold and platinum which gave standard values of the contact angle to within $\pm 4°$. The subsequent heat treatment consisted of maintaining the metal at a definite temperature for one hour. These temperatures ranged from 100 to 600° C.

The contact-angle measurements showed that both when water alone was present and when an organic liquid replaced the air, the angle measured in the water decreased considerably as the temperature of the treatment was raised. When the surface was subsequently exposed to air the contact angle gradually increased. In the case of the system solid-water-organic liquid the reduction of the water angle by heat treatment of the metal is expressed by the statement that the surface becomes less strongly organophilic and more strongly hydrophilic as the temperature of the treatment is raised.

The effect of the surrounding gas during heat treatment must be considerable and when this was investigated it was found that the metal surfaces became progressively less hydrophilic when heat treated *in vacuo*, in hydrogen, in air and in nitrogen in the order stated. This decrease was accompanied by an increase in their organophilic properties as instanced by their enhanced ability to react with organic compounds.

During the course of the work it is mentioned that Pyrex glass and brass when subjected to strain produced by compression or rapid cooling show increased contact angles with water. The effect disappears upon removal of the strain.

A critical study of the available methods of measuring the surface tension or surface energy of solids has been made by Saal and Blott[24]. Each method is examined, and all are rejected as liable to lead to incorrect results, the objections being made mainly on theoretical grounds. Such a paper cannot be fairly condensed in an article of this kind, but the present author thoroughly recommends it to the notice of those interested in the subject.

§ 5. THE SURFACE TENSION OF HEAVY WATER

In the *Progress Reports* for 1936 the results cited for the surface tension of heavy water were somewhat discordant. This year, however, better agreement seems to have been reached.

Measurements have been made on 99·7 per cent. D_2O by Timmermans and Bodson[25]. They first of all made careful determinations of the surface tension of H_2O by the methods of capillary rise, drop weight and bubble pressure at temperatures ranging from 0 to 40° C. The methods agree quite well and the values show a general falling off with rising temperature, but at 13° C. the curve shows a perfectly definite point of inflexion and becomes almost horizontal, falling again at the usual rate after a temperature of 15 or 16° C. is passed. This trend of the curve is somewhat novel, but each method of measurement confirms it. The surface tension of H_2O at 20° C. as indicated by their graph is about 72·9 dyne/cm. With regard to heavy water Timmermans and Bodson find that a curve of a similar shape is obtained, the point of inflexion occurring this time at 17° C. The result is that for most temperatures the surface tension of D_2O exceeds that of H_2O but between about 13 and 20° C. it is less. At these temperatures the values are equal and at 20° C. the 72·9 dynes/cm. given by the authors is said to agree with Flood and Tronstad[26] but not with Selwood and Frost[27].

Takéuchi, Sugita and Inai[28] have used a capillary U-tube method and find that over the range 10–90° C. the surface tension of D_2O is slightly below that of H_2O. From their curve σ_{D_2O} at 20° C. appears to be about 72·5 dyne/cm. The graph drawn assumes a more usual shape than does that of Timmermans and Bodson, although when the scatter of the points is closely examined it does seem possible that there is an indication of the latter authors' point of inflexion. An application of the Eötvös equation leads to a value of about 500° C. for the critical temperature of D_2O.

A third determination of σ_{D_2O} is that of Jones and Ray[29] who used their capillary method already mentioned in connexion with their work on electrolyte solutions. Measurements were made at 25° C. only, but great care was taken in the preparation of the specimens. Extrapolation to 100 per cent. purity gives $\sigma_{D_2O}/\sigma_{H_2O} = 0{\cdot}99946$, thus indicating a slight but (in the opinion of the authors) real difference between the two surface tensions.

§ 6. MISCELLANEOUS RESEARCHES

Theoretical. Ferguson and Kennedy[30] have made what must have been very laborious calculations relating to the free and total surface energies of liquids. They take the following equations as fundamental:

$$\sigma = \sigma_0 (1-m)^n$$

and

$$\sigma = C (\rho_{liq} - \rho_{vap})^p,$$

where σ_0 is the surface tension at absolute zero, m is the reduced absolute temperature and n can be found by graphical or other methods if sufficient surface tension-temperature data are available. If $p = 4$ the second equation becomes that of Macleod.

Next, following Katayama, the free molecular surface energy is defined as

$$e = \sigma \frac{M^{\frac{2}{3}}}{(\rho_{liq} - \rho_{vap})^{\frac{2}{3}}} = e_0 (1-m)^{n-\frac{2n}{3p}},$$

where $e_0 = \sigma_0^{1-\frac{2}{3p}} P^{\frac{2}{3}}$, P, the parachor, being defined quite generally as $MC^{\frac{1}{p}}$. Similarly the total molecular surface energy, E, is given by

$$E = \frac{\lambda M^{\frac{2}{3}}}{(\rho_{liq} - \rho_{vap})^{\frac{2}{3}}} = E_0 (1-m)^{\left(n-\frac{2n}{3p}-1\right)} \{1 + m(n-1)\},$$

where

$$E_0 = \sigma_0^{1-\frac{2}{3p}} P^{\frac{2}{3}} \quad (= e_0).$$

The total surface energy per sq. cm., λ, is related to the surface tension in the usual way by the Gibbs-Helmholtz equation.

An extensive table is given of values for σ_0, n, p and e_0 for a number of organic liquids and it is seen that n is usually slightly greater than 1·2 and p fluctuates about the value 4·0. It is precisely these fluctuations which give the calculations their interest; for instance, if the parachor is calculated from the appropriate value of p instead of on the assumption that $p = 4$, figures are obtained which, it is claimed, are even less dependent on temperature than are the more customary values.

Further, an equation of the type

$$\sigma_0 = K T_c^a V_c^b$$

is suggested, where T_c is critical temperature and V_c critical molecular volume. The authors find, on substituting the available data, that it can be written

$$\sigma_0 = 3{\cdot}12 T_c V_c^{-0{\cdot}55},$$

and that σ_0 calculated from this relation agrees well in all cases with the value calculated from surface-tension data. In a similar way it is found that

$$e_0 = 0{\cdot}811 T_c V_c^{-0{\cdot}176},$$

and if p is taken as being equal to 4, this leads to an expression for the parachor, viz:

$$P = 0{\cdot}324 T_c^{\frac{1}{4}} V_c^{\frac{3}{4}}.$$

Values of P based on this equation agree well with those evaluated by Sugden from surface-tension measurements.

Experimental. The detachment of drops in the drop-weight method of measuring surface tension has been photographed with a high-speed cinematograph camera at the rate of 1200 frames per sec. by Hauser, Edgerton, Holt and Cox[31]. The pictures reproduced show clearly the formation of the Plateau's spherule or secondary drop. This originates in the neck or stem which joins the falling drop to the liquid remaining on the jet. After rupture the neck becomes detached from both portions of liquid and falls separately showing the "nodes" characteristic of the initial stages of the breaking up of liquid cylinders. In the case of water, however, the neck does not disintegrate but occasionally a very small drop is ejected from the top of the secondary drop—sometimes with such violence that it bounces against the liquid still remaining on the jet. The fact that it does not coalesce with this liquid is ascribed to the action of electric charges. This possibility is also mentioned in connection with the behaviour of the secondary drop of nitrobenzene, which actually does break at the nodes into three separate drops which fall together for a short time and then spring apart. The benzene neck also disintegrates into a number of drops. When glycerol is used, however, its high viscosity causes the neck to begin as a thin filament, then to assume a dumb-bell shape with drops at each end which grow and rapidly approach each other until a single drop is formed.

For pure non-viscous liquids it is found that as the surface tension of the liquid decreases the length of the neck increases. The size of the secondary drop expressed as a percentage of the main drop also increases in the same way. If the surface tension of water is artificially lowered, however, by the addition of some surface-active substance the length of the neck becomes shorter than it is for pure water. This different behaviour from that of a pure liquid of low surface tension is possibly explained by the fact that the surface of a solution contains oriented molecules and has a different composition from its bulk.

The authors state that a deduction of the frequency of oscillation of the secondary drop from measurements made on the film leads to satisfactory values for the surface tensions of the liquids used.

Photographs of the detachment of a horizontal ring from a liquid surface were taken at the rate of 600 per sec. by the same authors. It is once again clearly apparent that it is mostly a matter of luck if the simple relation which equates the maximum pull to twice the surface tension multiplied by the mean circumference of the ring gives results anywhere near those yielded by other methods of measuring surface tension. After the detachment of the ring a film is formed in it, and this oscillates for a short time with a frequency which can be deduced from the photographs. Assuming that the surface tension is proportional to this frequency the authors calculate the surface tension of a liquid by comparison with water and reach the correct result.

Experiments of a somewhat different type have been carried out by Tsutsui[32]. He allowed a drop to fall on to a horizontal glass plate through which light was passing vertically and arranged a continuously moving film in such a way that the width of a dark strip on the film indicated the horizontal diameter of the drop after fall. The density of the darkening gave an idea of the thickness of the drop. The

photographic records show that the falling spherical drop spreads out rapidly on reaching the glass plate and shows a thin portion at the centre which disappears as equilibrium is reached. Impact velocity was varied and empirical relations have been established connecting the various quantities involved.

The question of dynamic and static surface tensions was the subject of some remarks by the present writer in the article on surface tension appearing in last year's *Progress Reports*. The position then seemed to be that methods of surface-tension measurement which involved rapid surface movements were apt to give high values as compared with the usually accepted ones even for pure liquids such as water and mercury. This year we note that Puls[33] has measured the surface tension of mercury by Bond's method in which a circular sheet of liquid is formed by the head-on collision of two jets. It is said that "the surface is renewed at the rate of about 100 times a second". He gives his most probable valueas $475{\cdot}5 \pm 2$ dyne/cm. which is well within the region in which (in the opinion of the present writer) the most reliable results now lie.

REFERENCES

(1) Fowler, R. H. *Proc. roy. Soc.* A, **159**, 229 (1937).
(2) Weng Wen-Po. *Phil. Mag.* **23**, 33 (1937).
(3) Bakker, G. *Z. phys. Chem.* **178** A, 193 (1937).
(4) Wagner. *Phys. Z.* **25**, 474 (1924).
(5) Onsager, L. and Samaras, N. N. T. *J. chem. Phys.* **2**, 528 (1934).
(6) Ariyama, K. *Bull. Chem. Soc. Japan*, **11**, 687 (1936).
(7) Ariyama, K. *Bull. Chem. Soc. Japan*, **12**, 32 (1937).
(8) Ariyama, K. *Bull. Chem. Soc. Japan*, **12**, 109 (1937).
(9) Jones, G. and Ray, W. A. *J. Amer. chem. Soc.* **59**, 187 (1937).
(10) McBain, J. W. and Wilson, D. A. *J. Amer. chem. Soc.* **58**, 379 (1936).
(11) McBain, J. W., Ford, T. F. and Wilson, D. A. *Kolloidzschr.* **78**, 1 (1937).
(12) Doss, K. S. G. *Proc. Indian Acad. Sci.* **4**, 97 (1936).
(13) McBain, J. W. and Swain, R. C. *Proc. roy. Soc.* A, **154**, 608 (1936).
(14) McBain, J. W. *Nature, Lond.*, **137**, 659 (1936).
(15) Merckel, J. H. C. *Kolloidbeihefte*, **45**, 413 (1937).
(16) Langmuir, I. *J. Amer. chem. Soc.* **39**, 1848 (1917).
(17) Lottermoser, A. and Winter H. *Kolloidzschr.* **66**, 276 (1934).
(18) Boas-Traube, Sonja and Volmer, M. *Z. phys. Chem.* A, **178**, 323 (1937).
(19) Wenzel, R. N. *Industr. and Engng Chem.* (Indust. Ed.), **28**, 988 (1936).
(20) Bartell, F. E. and Zuidema, H. H. *J. Amer. chem. Soc.* **58**, 1449 (1936).
(21) Bartell, F. E., Culbertson, J. L. and Miller, M. A. *J. phys. Chem.* **40**, 881 (1936).
(22) Bartell, F. E. and Miller, M. A. *J. phys. Chem.* **40**, 889 (1936).
(23) Bartell, F. E. and Miller, M. A. *J. phys. Chem.* **40**, 895 (1936).
(24) Saal, R. N. J. and Blott, J. F. T. *Physica*, **3**, 1099 (1936).
(25) Timmermans, J. and Bodson, H. *C.R. Acad. Sci.*, Paris, **204**, 1804 (1937).
(26) Flood, H. and Tronstad, L. *Z. phys. Chem.* A, **175**, 347 (1936).
(27) Selwood, P. W. and Frost, A. A. *J. Amer. chem. Soc.* **55**, 4335 (1933).
(28) Takéuchi, T., Sugita, T. and Inai, T. *Proc. Phys. Math. Soc. Japan*, **19**, 552 (1937).
(29) Jones, G. and Ray, W. A. *J. chem. Phys.* **5**, 505 (1937).
(30) Ferguson, A. and Kennedy, S. J. *Trans. Faraday Soc.* **32**, 1474 (1936).
(31) Hauser, E. A., Edgerton, H. E., Holt, B. M. and Cox, J. T. Jr. *J. phys. Chem.* **40**, 973 (1936).
(32) Tsuitsui, T. *Sci. Pap. Inst. Phys. Chem. Res., Tokyo*, No. 662, 227 (1936).
(33) Puls, H. O. *Phil. Mag.* **22**, 970 (1936).

SOUND

By E. G. RICHARDSON, B.A., Ph.D., D.Sc.

§ 1. FREE PROPAGATION: MISCELLANEOUS SUPERSONIC EFFECTS

Though it has always been held as a proof of the constancy of the velocity of sounds at all frequencies within the audible range that the sounds of a band should not be modulated—for better of worse—by distance, and although it has been pointed out that a more convincing demonstration to the unmusical ear could be sought in the comprehension of an open-air speech by a listener at the back of the audience, a scientific test of the matter seems until recently to have been lacking. Of course there have been laboratory measurements of the velocity of sound in tubes to fall back upon, but Walch[1] has recently given a convincing demonstration of the conservation of the wave form of a two-component note as it is radiated from the source—a loud-speaker—on to the detector—a collodion membrane at the apex of an exponential horn. The membrane was sputtered with silver, and an optical system gave by reflexion from it a spot of light which moved to and fro on a film in a high-speed drum camera. Records at a film speed of 32 ft./sec. were made of a fundamental + harmonic at distances of 75 and 250 cm. from the source. On these records the phase differences between the two components are compared, and it is concluded that between 250 and 1000 c./sec. the velocity cannot differ by more than 0·1 per cent.

Kuntze[2], by plotting the sound field round an oscillating quartz in air at various pressures, has shown that the absorption of sound (frequency 33,450 vib./sec.) in air is independent of pressure from 38 to 152 cm. of mercury, so confirming another canon in the theory of sound.

We have frequently mentioned in previous *Reports* the method of studying the velocity and absorption of supersonics in liquids, using the diffraction of light by the supersonic radiation. Argument still goes on about the theory of this effect and has now passed out of the realm of sound into that of optical theory, so that we shall not discuss it further here. We must, however, remark that attempts are now being made to detect diffraction of the light crossing a supersonic beam in a *gas*, and Bär[3] claims so to have measured the velocity of sound in air. The change of refractive index of air with pressure is not, however, large enough to make the diffraction spectrum easy to photograph, and the reproductions at present shown are not impressive.

Dorsey[4] gives an account of recent work performed by ships of the U.S. Coast and Geodetic Survey on the transmission of sound through sea water. Such work is fascinating in its complexity. In the words of the author: "Imagine a subaqueous laboratory, having two sides and one end open, the other end being a sloping wall, the floor usually irregular, the ceiling being a plane surface only for

long waves. The medium produces peculiar effects of refraction, reflection and diffraction. The experimenters may be from 10 to 50 km. apart." The present paper discusses the records of subaqueous signalling between the ships *Oceanographer* and *Lydonia* off the coast of Virginia in 1934. The records show that the direct signal is rapidly attenuated and fades within 10 km., but that the sound travels to much greater distances by multiple reflexions between the free surface and the sea-bed, confirming a hypothesis first made by Service.

Work on the coagulation by supersonics of aerosols, i.e. suspensions of solid or liquid particles in gases, to which we drew attention in the last *Report*, is being continued.* The workers at the Cologne Institute for Electrolyte Research (Brandt, Freund and Hiedemann[5, 6, 9]) have published more detailed results on this subject in which the rate of coagulation is determined not only by a series of micro-photographs at short time intervals, but also by a nephelometric method. This consists in allowing a beam of light to pass through the vessel containing the aerosol and to fall on a photoelectric cell. The extinction of the light is a function of the number N and average radius r of the particles, so that as they diminish in number faster than they gain in diameter by coalescing, the rate of coagulation is indicated by the rise in the photoelectric current. Both these authors and Andrade[7] have given theories of the effect based on König's work on the behaviour of particles in a sound field of frequency f, but whereas Brandt and Hiedemann attribute the action to coalescence of particles of different sizes due to their different mobilities (proportional to $r^2 f$ in their formula), Andrade first derives an expression for the attraction of neighbouring particles on the assumption that they remain stationary in the oscillating gas, and then introduces a correcting term for the mobility of the smaller and lighter particles. For the immobile particles he obtains the expression

$$N_0^{\frac{1}{2}} - N^{\frac{1}{2}} = kt,$$

where N_0 is the initial number of particles and N that after a time t, k being a constant independent of frequency (since mobility has been neglected). Parker[8], counting the number of particles of magnesium oxide in a graticule under the ultra-microscope at various times up to 8 min. from the commencement of the irradiation, obtained good agreement with this cube-root law (at a single frequency of 220 kc./sec.). The experimental work bristles with difficulties. There are so many effects—possible coagulation by electrical potentials, possible re-emulsifying of the coagulated aggregates, etc.—and there is such a large range of particle size in a polydisperse system like most practical aerosols, that it seems doubtful whether any comprehensive theory can be evolved. One must be content with a qualitative explanation of experiments in which the conditions are made as simple and fundamental as possible.

The dust figures formed by a pair of spark gaps, which when a discharge passes in series give rise to a short train of waves, have been used by Bate[10] to demonstrate interference phenomena. The wave-length was 1·7 cm., so that the radiation

* The effect was first noted by Patterson and Cawood in 1931.

bordered on the supersonic. The author used the method to trace the propagation of sound waves along a model horn section, and suggests that it may supplement the ripple tank in the study of acoustics of buildings. Tyler[11] too has demonstrated in such a tank a number of interference phenomena of which remarkably clear and brilliant projections or photographs may be obtained. The image is formed by reflexion from the surface of water or mercury, and a feature of the apparatus is the source of light which is a discharge tube synchronized to the reeds giving rise to the ripples.

§ 2. VIBRATING SYSTEMS: SOLIDS

Swigart[12] has devised a new electrodynamic method for maintaining a rod in longitudinal vibration, and used the apparatus to determine the velocity of sound in the material of the rod. The chief objection to mechanical methods of exciting longitudinal vibrations, such as that used in the elementary laboratory for Kundt's experiment, is that they tend to give the rod transverse vibrations which may modulate the natural longitudinal modes. Swigart overcomes this by winding a coaxial coil on the end of the rod and placing this in a radial magnetic field, so that all forces transmitted to the rod are directed along its axis. Close to the coil, the needle of an electro-magnetic pick-up is placed in contact with the rod, and the amplified e.m.f. generated in the armature of this drives the rod. The frequency is measured by a neon flash lamp and stroboscope, while the nodes are explored by a wandering pick-up connected through another amplifier to a pair of head phones.

Apart from papers of purely theoretical interest the only work on strings to report is connected with the violin and the pianoforte. Backhaus[13], whose work on the violin is well known, has recorded the acoustic and mechanical spectra of violins, mechanically played while preserving a constant bowing pressure. He varies the frequency by stopping the string and measures both the acoustic output at a fixed distance and the amplitude of the belly. Provided the load on the bow exceeds 25 g. weight the quality is found to be independent of bowing pressure. The acoustic and mechanical spectra go hand in hand. Most interesting are his observations at frequencies near the "wolf note". This is a beat between two coupled vibrations, that of the string, and one of those proper to the wood. As the tuning of the string is varied in the neighbourhood of the wolf note, the system behaves like a heterodyne oscillator in which the condenser of one of the coupled circuits is tuned while the other is unmoved.

Raisky[14] has devised a new stringed instrument, the "violon-basse", designed to play the bass part in amateur chamber music in the absence of a 'cellist. The instrument is of the shape and size of a viola (and so can be held like a violin) except for a greater depth to the belly giving a fundamental air resonance lower in the bass. In order to get comparatively short strings sounding at low pitch, covered strings are used, that is to say, there is a metallic core which bears the tension, and this core is wrapped round spirally with another material to increase the mass per unit length. The new instrument thus bears a similar relationship to the 'cello that the miniature pianoforte has to the full-scale model.

Vierling[15] has made a scientific study of the electrochord, an instrument which is like the common pianoforte as far as the production of tone is concerned but differs from it in having no sound-board. Instead, the vibrations of the strings act inductively on an amplifier and the sounds are given out by a loud-speaker, whose characteristics supply the formant usually added by the natural frequencies of the sound-board. Either piano timbre, with large damping, or organ timbre, with small damping, can be imitated. The latter is secured by a device which maintains the output of the amplifier at the maximum value reached just after the fall of the hammer, until the key is released. Of course, the musical characteristics of any electrical instrument such as this or the Hammond organ, which uses electromagnetic induction as the source of its notes, will be limited by the fidelity of the amplifier and loud-speaker.

Aoki[16] has recorded the sound of some bells hung in Japanese temples. These differ from the usual European type in that the thickness of metal continuously increases from the point of suspension to the mouth. The ratio of height to mouth diameter is about 3 : 2. Analysis of the wave-form showed that the lowest frequencies were inversely as the diameters of the mouths, while an empirical relation between the overtone frequencies and the fundamental is given, though without any theoretical backing.

Gramont and Beretski[17] have realized a new oscillator by placing together two quartz laminae separated by a thin film of balsam, and maintained in flexural oscillation by piezo-electric means. If we apply the usual formula for a bar in transverse vibration, we find that the fundamental frequency f is related to the thickness d and the length l by the formulae

$$f_1 = 8{\cdot}8 \times 10^4 d/l^2 \quad \text{or} \quad f_2 = 5{\cdot}66 \times 10^5 d/l^2,$$

according as the quartz is free or clamped. They envisage the possibility of quartz oscillators of quite low frequency. Thus a clamped pair 4 cm. long and 0·5 mm. thick has a fundamental of 275 c./sec. in transverse vibration. They have in fact made a number of these oscillators covering the gamut from 50 to 30,000 c./sec. The oscillations may be maintained by putting the quartz in parallel with a flashing neon lamp circuit.*

Another new oscillator due to Gruetzmacher[18] employs the usual longitudinal oscillations of the quartz, but the radiation from one face into a bath of oil is made unidirectional by enclosing an air cavity behind the other face. This cavity is formed in a small vessel fixed to the back of the quartz and of about the same size as the oscillator itself. These piezo-electric oscillators are now used so frequently to radiate energy into a liquid that it becomes important to study the effect of the liquid load upon them. A number of investigators have tackled this question. Parthasarathy[19] has obtained resonance curves of a quartz oscillator immersed in a liquid. Lowering of resonance frequency and broadening of resonance curves were observed, and hence the damping coefficients (in terms of air as standard) were calculated and compared with the relative viscosities of the liquids. Becker[20],

* A small "piezo-electric piano" of such oscillators is on view at the Paris Exhibition, 1937.

on the other hand, observed a *rise* in resonance frequency when the oscillator was immersed in the liquid, and ascribed this to inharmonic forced oscillations of the type which arise when the restoring force is proportional to some high power of the displacement. Further work, however, showed that a marked change in the decrement was produced in his crystal if the lower electrode were changed, which leads one to suppose that his curious observations were caused by change of pressure between crystal and electrodes when the liquid was introduced. Bär[21] has investigated the diffraction of the radiation in the liquid as it leaves the quartz and has connected this with the form of vibration of the quartz as exhibited by Chladni figures.

§ 3. VIBRATING SYSTEMS: FLUIDS

In an attempt to idealize the problem of the noise produced by airscrews and propellers—a problem of considerable industrial significance at the present time—Stowell and Deming[22] have taken a cylindrical rod, whirled it round an axle piercing the centre of the rod at right angles to its axis, and taken acoustic spectra of the sounds produced at various speeds. The aeolian tones given by the vortex systems so set up will cover a range of frequencies, for although the diameter of the obstacle is constant along its section the speed through the air is not. In fact a series of sounds is produced whose upper pitch limit is that of the aeolian tone proper to the rod if it moved bodily through the air at the tip-speed, figure 1. The output of acoustic power increases along the rod, as would be expected, except near the tips where it falls off a little, presumably due to "leakage" of vortices over the tips of the rods.

The fruitful field of edge-tone research has been invaded again by Krüger—who did pioneer work on this subject 25 years ago—in conjunction with Kaspar[23]. They give a critical survey of the whole phenomenon in the light of more recent research, and adduce in evidence of the similarity between edge-tone and jet-tone vortices some photographs which they have taken in a high-speed camera of smoke jets impinging on an edge. These they compare with the sensitive flame photographs of Brown (*Progress Reports*, **3**, 115). The arguments for and against Krüger's edge-tone theory would take too much of our space if elaborated here. Trendelenburg, Thienhaus and Franz[24] have made oscillograms of the establishment of organ-pipe sounds as the wind is turned on to the pipe, employing for this purpose filters which pick out different parts of the spectrum. In flue pipes there is a considerable change of quality in the first twentieth of a second as the tone passes through the underblown to the normal conditions of excitation. (These changes could in fact be explained in terms of edge tones, although they have not attempted this.) A reed pipe, however, comes directly on speech without noticeable change of quality. Particularly interesting are their records of the *vox humana*, a stop formed of pairs of reed pipes about a semi-tone apart which is intended to imitate the human voice, but has also been likened by a famous musician to the bleating of a goat. In fact, if one compares the *vox humana* records with those of a vowel such

as *ah* there is a similarity in the rapid dimunition of amplitude after the first peak in each period, but the substitute shows a greater preponderance of higher harmonics in its formant, probably due to the greater stiffness of the reeds as compared with that of the vocal chords.

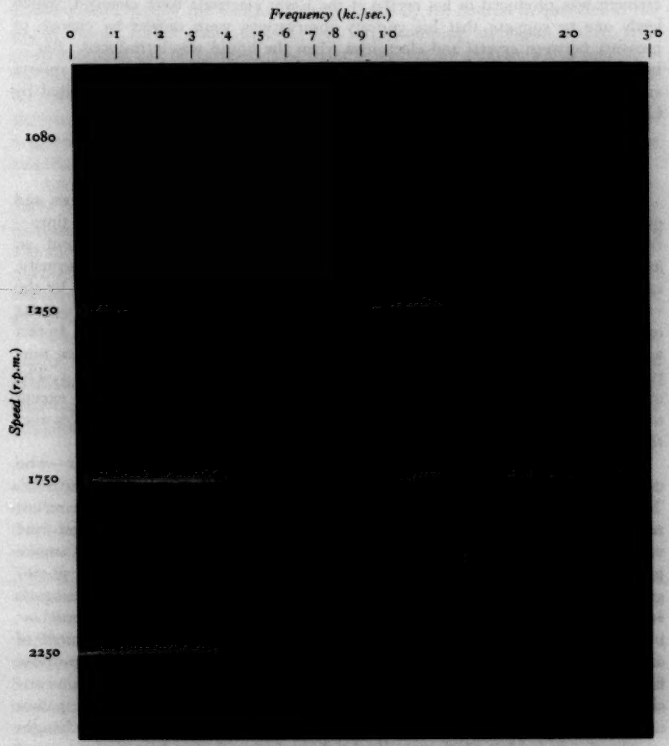


Figure 1

McLachlan and McKay[25] have worked out the theory of another type of transient, that which arises when a damped sinusoidal velocity is impressed on the throat of a loud-speaker horn. At high frequencies the horn acts as a pure resistance—pressure and velocity are in phase—and such components will be repro-

duced with little change of phase. The amplitude of components near the cut-off frequency, at which the horn becomes a pure reactance, will, however, be notably diminished due to the phase difference between pressure and velocity approaching $\frac{1}{4}\pi$. The matter is of practical import in the recording of rapidly damped sounds such as key-jingling, footsteps and paper rustling, which often seem too high pitched on sound films.

Jarnak[26] has also published (unfiltered) oscillograms of organ pipes and compared them with oscillograms of stringed instruments, while Bukofzer[27] has studied the acoustics of an instrument now only seen at Punch and Judy shows, viz. the Pan-pipe, a stopped pipe blown by an edge tone between the human mouth and one edge of the open end. He is concerned, however, not with the mechanism of the excitation but with the effect of different styles of blowing on the end correction, and hence on the fundamental of the pipe, still occasionally used by choral conductors as a pitch pipe.

King[28] gives us a more exact treatment of the end correction problem than the well-known one of Rayleigh. It may be remembered that the latter deduced a value of 0·824 for the ratio of the end correction to the radius of a pipe having an infinite flange. The actual problem was treated by analogy as that of the electrical resistance at the end of a cylindrical conductor terminating in an infinite medium. Actually Rayleigh used an approximation in which no provision was made for the continuity of the potential or its gradient on both sides of the end. King suggests an exact solution of the problem, making use of inversion theorems related to the theory of Bessel functions, which leads to a value for the ratio of 0·821, though the difference between this and Rayleigh's value will, as the author concedes, not amount to much in practice.

Two other pieces of research on wind instruments remain to be noticed. Young and Longridge[29] have passed a hot-wire probe through the barrel of a flute sounding with various combinations of finger stopping, with the object of finding to what extent the sounding column extends beyond the first open hole from the mouth. Aschoff[30] experiments with a mechanically blown clarinet. Here again we have a coupled system, reed plus air column, in which not inconsiderable changes in the natural frequency of the resonator may be produced by changes in the excitation of the reed, e.g. changes in the effective length of the reed due to the player's lips gripping it lower down, or in the damping due to a tight or loose grip. Particularly does reduction in the damping aid in overblowing, i.e. in producing the first overtone of the system. Stroboscopic observation of the reed shows that this exhibits high-frequency transients each time it is struck by the player's tongue at the beginning of a note, and these transients are held to be characteristic of clarinet quality.

We cannot close this section without mentioning a convenient loud source due to Hopwood[31] for demonstrating many phenomena of vibrating columns of air in tubes, striae and circulations in dust, formation and motion of liquid diaphragms in the tube, etc. The source consists of a large metal T piece with the long limb vertical and an Amal burner near its lower end. Intense sounds are produced on

the howling-tube principle, and the glass tube containing the dust or liquid is then pushed into the horizontal limb and the vibrations so communicated to it.

§ 4. ACOUSTIC IMPEDANCE: PROPERTIES OF MATERIALS

Lindsay and Barnes[32] discuss the theory of yet another type of acoustic filter, that of a solid rod loaded at equal intervals with flexible membranes which are threaded on to it like beads on a string. The rod is excited in longitudinal or torsional vibration and gives rise to transverse or radial torsional oscillations respectively in the membranes. Experimental confirmation of the results is at present lacking.

In investigations on the sensation of pitch it is very desirable to have the use of a pure tone of about 10 to 30 c./sec. Bekesy[33] has addressed himself to the construction of a source capable of producing simple harmonic air vibrations of low frequency.

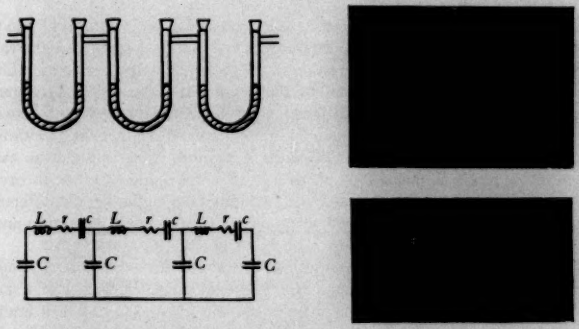


Figure 2

He first describes an ingenious filter to remove the overtones from the oscillations generated in a cylinder by a reciprocating piston. This consisted in essence of a long capillary tube containing three mercury pellets; actually the most convenient form was found to be three U-tubes, of bore 1·8 mm., containing mercury and forming a three-element low-pass filter. The arrangement is sketched in figure 2 with the electrical analogue below it. The capacitance, C, i.e. that of the air enclosed between the pellets, is

$$l_2/\rho c^2 S,$$

where l_2 is the length and S the cross-section of tube; while ρ is the density and c the velocity of sound in air, and the inertance L of the pellets (of length l_1 and density ρ_1) is $\rho_1 l_1 S$. The cut-off frequency

$$f_0 = \frac{1}{\pi\sqrt{(LC)}} = \frac{100}{\sqrt{(l_1 l_2)}}, \text{ approximately.}$$

Actually the cut-off is modified somewhat by the weight and viscosity of the

mercury, which are equivalent to a capacitance c, and resistance r respectively (cf. the circuit diagram). The oscillograms on the right of the figure show the wave-form of a 7 c./sec. oscillation before and after passing through the filter. Here $l_1 = 20{\cdot}2$ cm. and $l_2 = 9{\cdot}6$ cm.

After considering the possibility of using a thermophone in place of the reciprocating pump, the same author (to whom we owe many precision instruments for hearing testing) then goes on to describe a bridge method of comparing acoustic impedances, of which one arm is to be formed by the impedance of the outer ear. Schuster[34], too, now gives us more details of his acoustic bridge (cf. *Progress Reports*, **2**, 178). Whereas Bekesy uses capillary tubes for the ratio arms of his bridge, and assumes Poiseuille's Law to calculate their resistances, Schuster proposes for a norm a wide-flanged tube having an orifice at one end and a movable stopper at the other. As the stopper is moved over a distance of half a wave-length the impedance of the orifice goes through a cycle of values round about that which it would have if it were completely absorbent, i.e. if there were no reflexions from the backing tube. The smaller the cross-section (a) of the orifice with respect to that (A) of the flange, the less is the cyclic variation. In any case, the mean of the measured impedances of the orifice for a number of equidistant settings of the stopper is taken as the normal for purposes of comparison. Actually it is the absorption coefficient of small specimens of material which Schuster is concerned to measure, and he compares this with the theoretical absorption of the normal (in the absence of reflexion) which is

$$\frac{4Aa}{(A+a)^2}.$$

In this way he has tested absorbent specimens consisting of felt discs in a pile, bundles of drinking straws and German-sausage skewers (hollow wooden sticks) respectively.

Keibs[35] has a simple method for measuring the impedance, and hence the absorption coefficient, of a reflector at the end of a tube of length l with attenuation constant α and propagation constant β, provided βl is not less than 3 cot $(\alpha + i\beta)\, l$. He measures the pressure at the sending end when the other is closed successively by a hard stop and by the unknown impedance z. The ratio of these two pressures

$$p_1/p_2 = \frac{z}{z + \cot(\alpha + i\beta)\, l} = \frac{z}{z+1}, \text{ approximately.}$$

He uses this simple formula to measure the impedance of the ear drum placed at the end of a short tube, a microphone at the other end measuring pressures, while just behind it is a thermophone wire as source.

Of other absorption measurements we may mention those of Rankin[36], who has built a small reverberation and transmission chamber with panel $3 \times 2 \times 2$ ft. completely filling the partition between, and so obtains rather higher values than those using full-size chambers; and those of Cremer[37] which were done in the open air and embraced a range of incidence angles.

§ 5. SUBJECTIVE ASPECTS

Ferdinand Trendelenburg and Franz[38] have continued their analysis of speech sounds using filtration, octave by octave, in conjunction with an oscillograph. In the present paper they are principally concerned with the explosive consonants *p*, *t*, *k* and *b*, *d*, *s*. In producing the former unvoiced class of sounds the vocal cords are held open and the consonant noise is ascribed to jet tones produced in the throat and mouth as the breath is expelled. The bands are then brought together for the production of the following vowel. The time taken by the vortex noises which make up the consonant varies a good deal from one person to another but is of the order of a tenth of a second. In producing a vowel of the latter voiced group the vocal cords are first approximated and then set vibrating on the vowel with explosive force. Consequently the time to establish the vowel is much shorter, about one-hundredth of a second.

Wilhelm Trendelenburg[39] (brother of the author just mentioned) and also Lewis[40] have analysed a number of oscillograms from the point of view of the coupling between the vocal-cord vibrations and the formant introduced by the resonators of the mouth and pharynx. In particular, the former has given us what are probably the first simultaneous motion pictures of the opening and closing of the vocal cords in relation to the wave-form of the sound they produce, using an excised larynx of a dead calf. As we should expect, the vocal cords are open for a fraction only (about one-sixth) of each period, giving rise to strongly damped wave-forms. These have to be modulated in life by the resonators of the mouth cavity. Of all the common animals, the calf performs this modulation most efficiently; next to man, it has the most mellow voice.

Both Vierling[41] and Wagner[42] have carried out the synthesis of these formants by electrical circuits, initiated by Stewart in America in 1922. Vierling shows oscillograms of an electric wave of fundamental 200 c./sec. and rich in harmonics of almost equal amplitude; figure 3*a* shows its spectrum. He then passes this through a resonant circuit of 1100 c./sec. falling between the 5th and 6th harmonic; figure 3*b* shows, as continuous and broken lines respectively, the response and phase curves of this circuit. The effect of this is to superpose a number of the partials shown in (*a*) having amplitudes and relative phases as in (*b*) to give the curve of 3*c*. The experimental oscillogram with this simple formant agreed remarkably well with theory; in fact that of 3*c* recalls the wave-form of certain vowels, particularly *ah*.

Finally, Marro[43] has measured the extent to which speech may be broken by regular interruptions and still remain intelligible. About 10 interruptions per second of equal duration to the non-interrupted epochs is about the lowest value for intelligible listening, but even so aural flicker is perceived. At high frequencies of interruption there is a loss of low-frequency components in the speech. The optimum rate seems to be about 35 times per second. The matter is of importance in connexion with those telephone circuits for which sending and receiving is by common line or antennae, so that these must be connected rapidly in turn to the transmitting and receiving apparatus.

With the object of further testing the resonance theory of hearing, Hartridge[44] has repeated an old experiment of his. The idea is as follows. If the phase of a pure tone is suddenly changed by π, then if perception is by a resonator of some kind in the ear, this must cease vibrating for a moment before picking up again under the new conditions. A period of silence must therefore ensue when the phase is changed. Two main problems in performing this difficult experiment are (1) to ensure that the phase change shall take place neatly in the source without other superfluous noises (this rules out electric switches), and (2) to be sure that it *is* a period of silence that one hears and not a click or hiss. In spite of the precautions taken by Hartridge, the result of the experiment does not yet seem to be conclusive, and we await with interest further improvements in technique which are contemplated.

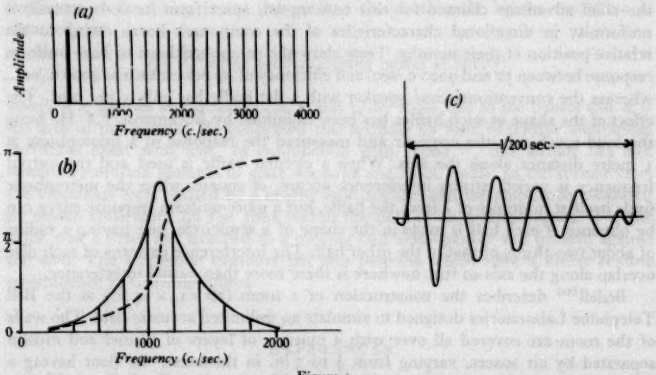


Figure 3

Further work has been done on the subjective change of pitch associated with a change of loudness, to which we referred in the last *Report* (p. 122). Snow[43] has confirmed that at low frequencies the pitch of a note appears to fall when the loudness is increased, but Lewis and Cowan[46] are unable to confirm this conclusion in their tests of the accuracy of tuning which string players can produce when they excite the strings with greater or less stress. There are a number of reports of tests on various types of defective hearing which have a physiological rather than a physical interest, but attention should be directed to an account by Littler[47] of the hearing aids in common use. This gives useful information with regard to the circuits and amplification of these instruments, not elsewhere available in collected form.

6. TECHNOLOGY

Of improvements in reproducing equipment, we note that Olson and Massa[48] have developed a compound horn, consisting of a diaphragm unit, a corrugated cone 5 in. in diameter, feeding on one side into a short horn with a small throat for the high-frequency range and on the other into a long horn with wide throat to bring out the low frequencies. The whole apparatus is placed in a box measuring $40 \times 40 \times 17$ in. without a front. The short horn comes straight forward from back to front, but the rear of the cone communicates with the throat of the large horn which is folded to and fro in the way now common in loud-speaker cabinets until it eventually debouches into an annular mouth surrounding the smaller one. Cinema loud-speaker units now frequently employ pairs of horns for the sake of fidelity, but the chief advantage claimed for this new model, apart from its compactness, is uniformity in directional characteristics of the component horns owing to the relative position of their mouths. Tests show the compound horn to have uniform response between 50 and 9000 c./sec. and efficiency of 40 per cent up to 2000 c./sec., whereas the conventional cone speaker with a flat baffle has only 5 per cent. The effect of the shape of such baffles has been examined by Buchmann[49].* He hung the loud-speaker in the open-air and measured the response of a microphone at 1 metre distance along the axis. When a circular baffle is used and the output frequency is varied, strong interference occurs, of course, when the microphone finds itself at multiples of λ from the baffle, but a more uniform response curve can be obtained if each half is made in the shape of a semicircle, one having a radius of about two-thirds of that of the other half. The interference patterns of each disc overlap along the axis so that nowhere is there more than partial interference.

Bedell[50] describes the construction of a room ($20 \times 14 \times 14$ ft.) at the Bell Telephone Laboratories designed to simulate an unlimited acoustic field. The walls of the room are covered all over with a number of layers of flannel and muslin separated by air spaces, varying from $\frac{1}{2}$ to 3 in. in thickness, the floor having a grating supported free of the absorbent for the experimenters to walk on. Measurements of the absorption coefficient of the material both by tube methods and in the complete room showed this to average 97 per cent above 150 c./sec. Such rooms are, of course, of great interest in measurements of audition.

Meyer[51] has tested the frequency characteristics of this type of absorbent, more particularly in the case where the membrane enclosing the air space is not porous. In such a case the natural damped vibrations of the air cell play a large part in its sound absorbing qualities. The electrical analogue of this system is a resistance in series with an inductance and a condenser. The mass m of the membrane fulfils the function of the inductance; the capacitance is that of the cavity and $= l/\rho c^2$, where l is the depth, supposed small compared to the wave-length of the sound. The stiffness of the membrane contributes to the resistance. The natural (fundamental)

* Readers interested in history may care to note, as this author reminds us, that baffles in the sense that they are used on loud-speakers date back to Stokes. He put them round tuning-fork prongs.

frequency is then $(c/2\pi)\sqrt{(p/ml)}$ when the resistance is neglected. The absorption is naturally very selective. Damping of the air cavity by introduction of cotton-wool increases the absorption but makes the resonance peak sharper; this effect the author ascribes to coupling between membrane and air cell. He suggests that such a system, e.g. a paper membrane 5 cm. from a wall, with the interspace stuffed with cotton-wool, would be useful in technical acoustics when it is desired to absorb the lower frequencies at the expense of the higher.

Constable[52] has also considered the effect of an absorbent lining between double partitions both of which may drum. His theoretical treatment is more general than Meyer's, embracing as it does the ricochetting of the sound energy within the cavity between the two panels, some being absorbed by the lining and another fraction being transmitted through the panel at each rebound. If the absorption coefficient of the lining is not too great, a simple formula for the insulation is obtained, viz.

$$R=\frac{r_1r_2}{A}\times(\text{the total absorbing power of cavity}),$$

where R is the net sound reduction, r_1 and r_2 are those due to the panels and A is the area of the panels. He confirmed this formula by tests of double aluminium partitions lined with felt. In conjunction with Aston[53], this author has also obtained vibration patterns of glass windows and brick walls. In the former case, the vibrations were excited by a nearby loud-speaker and the amplitude measured by search coils stuck on the glass within the field of an electromagnet. For the brick wall a brass rod attached to a moving coil loud-speaker was pressed against the wall. The amplitude contours fall into patterns which have a rather remote resemblance to Chladni figures.

REFERENCES

(1) Walch, F. A. *Proc. phys. Soc.* **48**, 899 (1936).
(2) Kuntze, A. *Ann. Phys., Lpz.*, **26**, 349 (1936).
(3) Bär, R. *Helv. phys. Acta*, **9**, 367 (1936).
(4) Dorsey, H. G. *J. acoust. Soc. Amer.* **7**, 287 (1936).
(5) Brandt, O. and Hiedemann, E. *Kolloidzschr.* **75**, 129; **76**, 272 (1936).
(6) Brandt, O. and Hiedemann, E. *Trans. Faraday Soc.* **32**, 1101 (1936).
(7) Andrade, E. N. da C. *Trans. Faraday Soc.* **32**, 1111 (1936).
(8) Parker, R. C. *Trans. Faraday Soc.* **32**, 1115 (1936).
(9) Brandt, O., Freund, H. and Hiedemann, E. *Kolloidzschr.* **77**, 103 and 168 (1936).
(10) Bate, A. E. *Proc. phys. Soc.* **48**, 178 (1936).
(11) Tyler, E. *Proc. phys. Soc.* **48**, 659 (1936).
(12) Swigart, J. I. *Rev. sci. Instrum.* **7**, 252 (1936).
(13) Backhaus, H. *Z. tech. Phys.* **17**, 573 (1936).
(14) Raisky, D. *C. R. Acad. Sci., Paris*, **203**, 1137 (1936).
(15) Vierling, O. *Z. Ver. dtsch. Ing.* **80**, 1069 (1936).
(16) Aoki, I. *Mem. Coll. Sci. Kyoto*, **19**, 283 (1936).
(17) Gramont, A. de and Beretski, D. *C. R. Acad. Sci., Paris*, **202**, 1229 (1936).
(18) Gruetzmacher, J. *Z. tech. Phys.* **17**, 166 (1936).
(19) Parthasarathy, S. *Proc. Ind. Acad. Sci.* **3**, 544 (1936).

(20) Becker, H. E. *Ann. Phys., Lpz.*, **25**, 359; **26**, 645 (1936).
(21) Bär, R. *Helv. phys. Acta*, **9**, 617 (1936).
(22) Stowell, E. Z. and Deming, A. F. *J. acoust. Soc. Amer.* **7**, 190 (1936).
(23) Krüger, F. and Kaspar, H. *Z. tech. Phys.* **17**, 416 (1936); *Phys. Z.* **37**, 842 (1936).
(24) Trendelenburg, F., Thienhaus, F. and Franz, E. *Akust. Z.* **1**, 59 (1936).
(25) McLachlan, N. W. and McKay, A. T. *Camb. phil. Soc.* **37**, 265 (1936).
(26) Jarnak, P. *Fysisk Tidsskr.* **34**, 137 (1936).
(27) Bukofzer, M. *Z. Phys.* **99**, 643 (1936).
(28) King, L. V. *Phil. Mag.* **21**, 128 (1936).
(29) Young, R. W. and Longridge, D. H. *J. acoust. Soc. Amer.* **7**, 178 (1936).
(30) Aschoff, V. *Akust. Z.* **1**, 77 (1936).
(31) Hopwood, L. F. *Nature, Lond.*, **138**, 1059 (1936).
(32) Lindsay, R. B. and Barnes, T. C. *J. acoust. Soc. Amer.* **8**, 42 (1936).
(33) Bekesy, G. v. *Ann. Phys., Lpz.*, **25**, 413; **26**, 544 (1936).
(34) Schuster, K. *Elekt. Nachr.-Tech.* **13**, 164 (1936).
(35) Keibs, L. *Ann. Phys., Lpz.*, **26**, 585 (1936).
(36) Rankin, J. S. *J. roy. tech. Coll., Glasgow*, **3**, 542 (1936).
(37) Cremer, L. *Elekt. Nachr.-Tech.* **13**, 36 (1936).
(38) Trendelenburg, F. and Franz, E. *Wiss. Veröfft. d. Siemens. Konzern*, **15**, 78 (1936).
(39) Trendelenburg, W. *S.B. preuss. Akad. Wiss.* pp. 308 and 338 (1936).
(40) Lewis, D. *J. acoust. Soc. Amer.* **8**, 91 (1936).
(41) Vierling, O. *Ann. Phys., Lpz.*, **26**, 219 (1936).
(42) Wagner, K. W. *S.B. preuss. Akad. Wiss.* p. 2 (1936).
(43) Marro, M. *Phil. Mag.* **22**, 847 (1936).
(44) Hartridge, H. *Proc. phys. Soc.* **48**, 145 (1936).
(45) Snow, W. B. *J. acoust. Soc. Amer.* **8**, 14 (1936).
(46) Lewis, D. and Cowan, M. *J. acoust. Soc. Amer.* **8**, 20 (1936).
(47) Littler, T. S. *J. sci. Instrum.* **13**, 144 (1936).
(48) Olson, H. F. and Massa, F. *J. acoust. Soc. Amer.* **8**, 48 (1936).
(49) Buchmann. *Z. tech. Phys.* **17**, 563 (1936).
(50) Bedell, E. H. *J. acoust. Soc. Amer.* **8**, 118 (1936).
(51) Meyer, E. *Elekt. Nachr.-Tech.* **13**, 95 (1936).
(52) Constable, J. E. R. *Proc. phys. Soc.* **48**, 690 and 914 (1936).
(53) Constable, J. E. R. and Aston, G. H. *Proc. phys. Soc.* **48**, 919 (1936).

SUPERSONICS IN RELATION TO MOLECULAR CONSTITUTION

By E. G. RICHARDSON, B.A., Ph.D., D.Sc.

§ 1. INTRODUCTION

The measurement of the velocity of sound in gases has since the time of Kundt formed an important field of exploration for physicists and physical chemists, interested in the factor γ—the ratio of specific heats at constant pressure and constant volume—as an important indicator of the make-up of a molecule. It is only within the last few years, however, that the new science of supersonics has furnished evidence of a quite unexpected nature on the structure of gases and liquids, which, though in part referred to in earlier Reports, has reached a stage when a critical survey may with profit be attempted. It is not the intention to discuss the methods by which the experimental results are obtained; these are described in earlier Reports.

§ 2. OBSERVATIONS ON THE PROPAGATION OF SUPERSONICS IN GASES

The story begins with an observation of Pierce[1] in 1925, that the velocity of sound of frequency in the neighbourhood of 100 kc./sec. exceeded that at sonic frequencies, and further that the radiation was strongly attenuated in this region, making velocity measurements difficult or impossible at a moderate distance from the source. Since then the results have been extended, and the effect has been established for carbon dioxide, nitrous oxide (at 100 kc./sec.), carbon disulphide (at 500 kc./sec.), sulphur dioxide (at 2000 kc./sec.), and ammonia (at 300 kc. sec.). There is doubt whether it occurs in chlorine or acetic acid vapour[2]. Pure gases in which no difference between the low and high frequency measurements (up to 2000 kc./sec.) has been found include argon, hydrogen, oxygen and nitrogen. (Dispersion in air, which was formerly found with the dust-figure method and attributed to the last two of these gases, has recently been ascribed by Parker[3], when repeating the observations, to an error in measurement.) It should be noted, in view of the discussion later, that with one exception the gases showing dispersion of the velocity with excessive absorption are tri-atomic symmetrical molecules of the O—C—O type, while the "failures" include the most common diatomic molecules.

Of some importance to the theory of the effect are the measurements initiated by Richards and Reid[4] by a few measurements in CO_2 at various pressures and temperatures. These have been carried over a considerable range by Railston[5], Penman[6] and the writer in CO_2 (pressure and temperature), N_2O and SO_2 (pressure only), and by Steil[7] recently in NH_3 (pressure only). The pressure results fall nicely on a curve when plotted against the parameter frequency pressure

(figure 1)—note, however, the difference in vertical scales—but plotting the results of Richards, Reid and Penman for the temperature effect (after reducing velocities to 0° C. in the usual fashion) to get a single curve is not so simple, although the frequency corresponding to the middle of the dispersion rise (when $V = 262$ m./sec.) shows a linear increase with temperature, figure 2. (It is, of course, much easier to keep the apparatus at the same pressure all over than to preserve equality of temperature.)

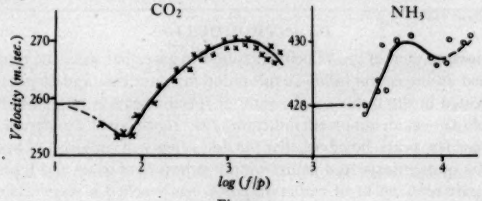


Figure 1

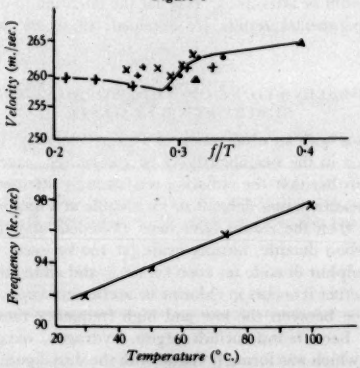


Figure 2

The symbols in the upper figure refer to different frequencies; those in the lower figure are for a velocity of 262 m./sec.

Another important effect was discovered by Knudsen[8]; this time in moist gases at comparatively low frequencies. He found that the absorption of sound in air was largely affected by the presence of water vapour. There was a maximum absorption at a concentration which varied with the frequency (between 3 and 10 kc./sec.) but averaged 0·4 per cent of water molecules. Since that discovery the effect on the absorption and velocity of sound in the acoustically inert gases by the admixture of a slight impurity has been much explored both in the sonic and supersonic gamut.

With the exception of water most of the impurities cause an upward shift δf of the frequency for maximum absorption in direct proportion to the concentration c of the adjunct. Thus

$$\delta f = ac,$$

but for water

$$\delta f = ac + bc^2,$$

where a and b are constants. The value of a varies greatly between different gases, and seems to be characteristic of the molecular structure, though its precise significance remains to be discovered. Some relative values of a are given in table 1; the gas with which the mixture is made is oxygen in every case.

Table 1

Impurity	Relative values of a
Ethyl alcohol	60
Ammonia	16
Benzene	13
Ethylene	4
Carbon tetrachloride	1

The diatomic gases and carbon dioxide showed a value of a less than 1. Although shift of the absorption peak is comparatively easy to observe, the recorded effects on velocity are small, and, one would imagine, difficult to measure. For example, in the year under review, Knudsen and Obert[9] have readily detected the shift of the peak absorption frequency in oxygen with a percentage of added water vapour or ammonia up to 3·5 kc./sec., whereas Richards[10] has to look for a change of < 1 per cent in the velocity in ethylene, when small quantities of deuterium are added. The effect of impurity on absorption is in fact well established in the case of oxygen at *sonic* frequencies with added water or ammonia. In cases involving *supersonic* dispersion, owing to the paucity of results as yet available, it is better to reserve judgement.

§ 3. OBSERVATIONS ON THE PROPAGATION OF SUPERSONICS IN LIQUIDS

In liquids again, it is the excessive absorption shown by certain organic compounds which is the salient phenomenon. It was Biquard[11] who first observed this in 1933, when he had developed the technique of diffraction of light by supersonics to the degree that he could employ an oscillator of 8000 kc./sec. as source. In table 2, his most recent values[12] for the absorption coefficient are shown in comparison with the values to be expected, if the absorption had been entirely due to viscous absorption (and if this had been governed by classical laws).

Table 2

Substance ...	Toluene	Xylene	Ether	Benzene	Chloro-form	Acetone	Water
Absorption coefficient × 10^3:							
Calc.	4·9	7·8	5·7	5·4	6·6	4·4	5
Obs.	55	47	35	580	300	20	15
Light-scattering powers	0·95	1·25	0·36	1·0	0·5	—	0·06

Bazulin[13] has also published results of the same order in acetic acid. Between 3730 and 31,300 kc. sec. he finds an approximately linear increase in the absorption coefficient from 0·4 to 1·85, passing close to the value 0·5 of Biquard for ethyl acetate at 8000 kc. sec. Apart from the enormous attenuation, a linear relation between absorption and frequency is itself in conflict with classical theory, which would make the absorption (due to viscosity and heat conduction alone) proportional to the square of the frequency. The same author has shown that the absorption coefficients of mixtures of these organic substances are not additive, a behaviour which recalls that of moist air in this respect. Sörensen[14] has examined water and a number of organic liquids between 200 and 900 kc./sec. He finds a slight decrease of absorption in water with rising frequency, although, as he points out, this liquid is more likely to be affected in its absorbing characteristics by air in solution than other liquids. Apart from this, the absorption coefficients in water and alcohol are of the same order as the predicted ones. In xylol, toluene, acetone and benzene however, he obtains values which tend to those of Biquard at higher frequencies. To this list we must add carbon disulphide, which bears out its anomalous behaviour in the gaseous state by excelling as liquid with an absorption about one thousand times normal! This result is due to Claes, Errara and Sack[15], who, however, find a notable diminution of absorption if small quantities of carbon tetrachloride are added to the bisulphide. Both Baumgardt[16] and Sörensen[17] have tried the effect of temperature on this absorption. Whereas the former at 800 kc./sec. in water reports the correct change with temperature which would have been expected from the known change of the coefficient of dynamical viscosity, but, of course, at a higher absolute level, the latter at 530 and 950 kc./sec. finds a rise of absorption with rising temperature in acetone, alcohol, toluene and xylol, whereas, of course, the viscosity drops. Alcohol and benzene did not show agreement between their behaviour at the two frequencies under investigation.

In spite of these large absorptions, variations in velocity with frequency are not easy to detect in liquids. Thus Parthasarathy[18], who has measured velocities in a large number of organic liquids with a view to obtaining a connexion between adiabatic compression and velocity, is unable to detect any appreciable difference in the velocity of sound in a number of hydrocarbons, although it is true that he only used two frequencies (7000 and 20,000 kc./sec.). On the other hand, Hiedemann, Seifen and Schreuer[19], who claim to be able to measure wave-lengths to 0·01 per cent and temperature to 0·01°, in their apparatus have detected a dispersion in toluene, xylol and aniline. In the preliminary notice of their work they mention only two frequencies, i.e. 4800 and 8400 kc./sec., so that it is not clear whether they have mapped out complete dispersion curves. Toluene and xylol show a rise of 1 in 1300 m./sec. between these two frequencies, whereas the velocity in water is unchanged. Liepmann's[20] results on liquid oxygen between -183 and $-209°$ C. at 7500 c./sec. should also be mentioned here, since they show a trend of the velocity with temperature which gives a curve of peculiar shape for the variation of adiabatic compressibility with temperature.

§ 4. DISCUSSION OF THEORIES OF SUPERSONIC DISPERSION

Dispersion of sound, i.e. changes of velocity with frequency and consequent attenuation of amplitude at a rate greater than can be accounted for by undisturbed radiation, can be traced to one or more causes, which we shall discuss in turn: (1) viscous damping, (2) relaxation of parts of the medium, (3) resonance of parts of the medium and (4) anistropy of the medium.

(1) *Viscosity* by itself can produce dispersion. If we suppose that the Helmholtz-Kirchhoff laws governing the decay of sound in a viscous medium still apply at these high frequencies—and we have indirect proof of this, as is described below—we should observe an absorption proportional to the square of the frequency. This is accompanied by a gradual fall in the velocity, so slight, however, as to render its detection difficult. Actually the absorption does behave in this way in gases like argon and oxygen, though the absolute value of the absorption coefficient is higher than that given by Kirchhoff, but, of course, nothing like so great as in carbon dioxide and others of that class. Evidence that viscous damping of supersonics behaves in the same fashion as that of low-frequency vibrations is afforded by the very recent work of May[21], who showed that the absorption of supersonics in air contained in capillary tubes (where the viscous drag of the walls enhances the absorption, and so makes it easier to measure than in the open air) gave results in keeping with that of low-frequency pulsations. In spite of the rapid attenuation of the supersonics in the tubes, the lowering of velocity was very small and uncertain.

(2) The relaxation theory began with Jeans[22], who in 1904 worked out the effect of a lag in the transfer of translational into vibrational and rotational energy in a system of loaded spheres on the absorption and velocity of a disturbance passing through them. He reckoned that for sound waves, the effect would be small compared with that of viscosity. Herzfeld and Rice[23] first applied the idea to explain Pierce's results in carbon dioxide and since then the theory has been elaborated and modified by a number of scientists. It will perhaps be best to describe the effect in the words of one of its exponents. Groszmann[24] writes:

"In a fluid which contains not only translational (external) energy but also rotational or vibrational (internal) energy, the number of molecules excited in rotation or vibration is a function of the temperature. In most cases a finite time is necessary to establish equilibrium between temperature and internal energy, and one which is peculiar to each rotational or vibrational degree of freedom. One can thus suppose that every molecular state, i.e. every quantum of rotation or vibration, takes a certain time to set up (usually known as 'relaxation time'). If for a particular quantum state this time is so large that it can only partially or not at all follow the adiabatic temperature changes in the sound wave, so that, therefore, when the temperature rises or falls no more energy can be taken or given up, then the specific heat becomes less as the frequency rises by an amount representing the energy quantum involved. The velocity of sound is accordingly raised.

"The absorption of sound is also affected. Since the energy of translation does

not at once take up its equilibrium value, the pressure which is proportional to this energy experiences a phase displacement relative to the oscillating density, and acoustic energy is turned irreversibly into heat."

The rise in the dispersion curve of a gas (cf. figures 1 and 2) occurs then, according to this theory, when the time-period of the sound approaches the mean relaxation time of the molecules. In this way Herzfeld and Rice predicted a value of 10^{-5} sec. for the relaxation time in CO_2. The same authors obtained an expression for the rise in velocity to be expected. This has been put in several forms by different authors. The treatment which follows is due to Henry[25].

Let E_x be the actual energy and E_T the energy which the molecules would have at equilibrium at temperature T, and suppose that

$$\frac{dE_x}{dt}=\frac{1}{\beta}(E_T-E_x) \qquad \ldots\ldots(1),$$

where β is the period of relaxation of the vibrational energy. If the gas is subject to adiabatic variations of temperature of frequency $\omega/2\pi$,

$$T=T_0+T_1e^{i\omega t},$$

$$E_T=E_0+C_\omega T_1e^{i\omega t},$$

C_ω being the specific heat of vibrations. Substituting in (1) and solving, we find

$$E_x=E_0+C_\omega T_1(1+i\omega\beta)^{-1}e^{i\omega t}.$$

If the total specific heat be regarded as made up of two parts, C_ω due to vibration and C_1 due to translation (+rotation),

$$C'=C_1+C_\omega(1+i\omega\beta)^{-1},$$

where C' is complex; also, as usual, $\gamma'=1+R/C'$, R being the gas constant, and the velocity of plane waves V is $\sqrt{(\gamma' p/\rho)}$. Hence

$$\gamma'=1+R[C_1+C_\omega/(1+i\omega\beta)]^{-1},$$

and is also complex.

To find the real part we write

$$\frac{1}{C_1+C_\omega/(1+i\omega\beta)}=\frac{1+i\omega\beta}{C_\omega+C_1+i\omega\beta C_1},$$

of which the real part is

$$\frac{C_0+\omega^2\beta^2C_1}{C_0^2+\omega^2\beta^2C_1^2} \qquad \ldots\ldots(2),$$

where $C_0=C_\omega+C_1$, i.e. the total specific heat when $\omega\to 0$. Thus the usual specific heat at constant volume in the expression for the velocity of sound is replaced by the reciprocal of (2). When ω approaches the value $C_0/C_1\beta$ a rise in velocity occurs from the normal $V_0^2=(1+R/C_0)\,p/\rho$ to an ultimate value $V_\infty^2=(1+R/C_1)\,p/\rho$.

The disperson so produced is similar to that of electromagnetic waves in polar liquids (Debye).

The relaxation time β is also linked with the probability that a given molecule will be excited in vibration or rotation. The smaller this time becomes the greater will be the number of excited molecules at any instant. On this theory, the effect of

impurities on relaxation time is to increase the number of excited molecules of the (originally) pure gas, and so increase the probability that a molecule will be in the excited state. Collisions with dipole molecules or molecules with which the original ones have a chemical affinity should be most effective in this sense (cf. table 1).

(3) *Selective resonance* of the type we have in mind would have the same effect as the anomalous dispersion of the velocity of light observed by Wood in sodium vapour irradiated by a source near its natural frequency of emission. This type of dispersion has been observed with low-frequency sound waves passing through a medium filled with resonators all tuned to the same frequency. The dispersion curve, figure 3*c*, has been worked out by Belikov[26] and follows closely the lines of the one familiar in optics and transmission engineering, wherever, in fact, there is selective resonance.

(4) A medium filled with a large number of obstacles, e.g. an irregular suspension of solid spheres in a liquid, or an emulsion of drops of one liquid in another, can disperse sound. This effect will be at a maximum when their separation approximates to the half wave-length, or an odd multiple of the half wave-length, of the sound in virtue of interference between waves reflected from front and rear.

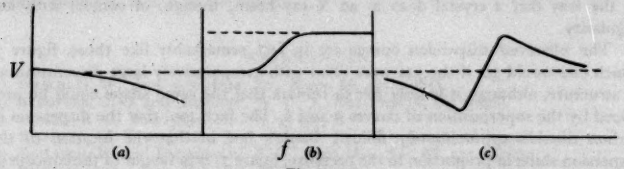


Figure 3

The dispersion caused by interference is of the same type as anomalous dispersion, and the writer does not know how the two can be distinguished except in degree. The three theoretical types of dispersion curves due to (*a*) friction, (*b*) relaxation and (*c*) selection are shown in figure 3. An important question is, What happens to the supersonic energy in a dispersive gas or liquid? The relaxation theory suggests that it should be turned into heat. Now a rise of temperature has certainly been observed in oily liquids when they are strongly irradiated, and has been ascribed to viscous damping, but although carbon dioxide exhibits a higher attenuation coefficient (value rising from 1 to 3 per cm. between 100 and 2000 kc./sec., as against 0·6 in benzene at 8000 kc./sec.) no rise of temperature has to the writer's knowledge been detected by resistance thermometers placed in front of a supersonic oscillator in carbon dioxide. What certainly does take place is marked *scattering* of the supersonic energy, so that what would be a straightforward beam in a normal fluid becomes a broadcast in the apparently absorbent fluid. This was first observed by the writer[27] in gaseous carbon dioxide in 1934 and has now also been detected by Biquard[28] in liquid toluene, acting on a suggestion of Lucas[29] who ranges himself whole-heartedly in favour of the anistropy principle, at any rate as far as liquids are concerned. It is indeed significant, as the writer has before remarked, that most of

the substances which show the scattering of sound waves behave in the same way to light. The two most noteworthy sound-scattering gases (CO_2, N_2O) stand highest in the list of light-scattering gases, while benzene and toluene head the list of vapours.* As liquids, benzene, toluene and xylene give high intensities in the Tyndall cone of scattered light. (If the reader turns back to table 1 he will see these optical constants, in respect to benzene as unity, added for comparison with the acoustical ones.) Again, benzene is a liquid in which the Raman effect is often demonstrated. Lucas, in fact, starting with the Lorentz theory of perturbations, derives a formula for the diffusion of sound by microscopic variations in the density, which differs but little from the corresponding formula of Rayleigh for light scattering. The observed value of the scattering, whether in carbon dioxide or toluene, far exceeds that which can be ascribed to microscopic fluctuations, so that it appears as though some form of aggregation was taking place. Mandelstam and Leontovic[30] have also drawn attention to these statistical fluctuations in the density of a liquid which can be evoked to explain, among other things, the doublets formed in the lines of scattered light. Such fluctuations will, in their opinion, give the medium a sort of crystalline structure able to diffract an incident beam of sound in the way that a crystal does to an X-ray beam, though, of course, with less regularity.

The observed dispersion curves are in fact remarkably like those, figure 3, which one would get if the gas were filled with comparatively large discontinuities in structure, although it is only fair to remark that the same shape could be produced by the superposition of curves *a* and *b*. The fact, too, that the dispersion in carbon dioxide approximately follows Boyle's law in that the location of the dispersion shifts in proportion to the pressure, figure 1, is in favour of the concept of a turbid medium in which the scattering centres expand and contract with surrounding conditions (cf. the effect of pressure and temperature on optical dispersion in gases). The same effects have been explained by the relaxationists by postulating changes in the excitation of the molecules engendered by pressure and temperature. These are, however, in the nature of *ad hoc* explanations. They do not appear to be based on fundamental principles or even on analogies with allied phenomena. Some indeed go so far as to invoke four or five vibration or rotation states in a molecule, basing their theory on the location of one or two points on the dispersion curve. Only a Rayleigh could do this with impunity, and until the technique of measurements above 1000 kc./sec. is further developed, to give more accurate velocity measurements, such niceness of detail in theory is to be deprecated.

One other point in favour of the anisotropic theory concerns the considerable change in dispersive power of a gas or liquid which can be produced by the admixture of a small impurity. By so doing we are obviously affecting the anisotropy to a much greater extent than a small change of pressure or temperature can induce. Lucas cites the case of a (liquid) mixture of nitro-benzene and hexane—proportions not stated—which had a supersonic absorption 30 times as great as that of

* The latest gas to be added to the dispersion list, ammonia, seems to be an exception.

either constituent alone, while we have already remarked the absorption which can be produced in air at lower frequencies by a little water vapour.

§ 5. CATALYTIC ACTION

We may conclude by referring briefly to another effect of supersonic radiation which may be of molecular origin, i.e. that in which the radiation acts as a catalyst to chemical reactions. We have, in fact, referred in previous *Reports* to experiments in which organic substances are reduced to simpler molecules in the presence of intense supersonic radiation. In view, however, of the lack of adequate control which seems to have been exercised in most of these assisted reactions, particularly the lack of assurance that the action is not mainly a physical one induced by the excessive vibration and heat generated, it would be premature to discuss these experiments further at the moment under the title of this *Report*. It is hoped, however, that what has been written here on the propagation of supersonics in gases and liquids, while in no sense professing to be a finished survey of a still progressing study, may encourage others, and particularly chemical physicists and physical chemists, to join in the search.

REFERENCES

(1) Pierce, G. W. *Proc. nat. Acad. Sci., Wash.* **60**, 271 (1925).
(2) Strother, C. O. and Richards, W. T. *J. chem. Phys.* **4**, 561 (1936).
(3) Parker, R. C. *Proc. phys. Soc.* **49**, 95 (1937).
(4) Richards, W. T. and Reid, J. A. *J. chem. Phys.* **2**, 193 and 206 (1934).
(5) Railston, W. and Richardson, E. G. *Proc. phys. Soc.* **47**, 533 (1935).
(6) Penman, H. L. *Proc. phys. Soc.* **47**, 543 (1935).
(7) Steil, O. *Z. phys. Chem.* **31**, 343 (1936).
(8) Knudsen, V. O. *J. acoust. Soc. Amer.* **3**, 126 (1931); **5**, 112 (1933).
(9) Knudsen, V. O. and Obert, L. *J. acoust. Soc. Amer.* **7**, 249 (1936).
(10) Richards, W. T. *J. chem. Phys.* **4**, 561 (1936).
(11) Biquard, P. *C. R. Acad. Sci., Paris*, **197**, 309 (1933).
(12) Biquard, P. *Ann. Phys., Paris*, **6**, 195 (1936).
(13) Bazulin, P. *C. R. Acad. Sci. U.R.S.S.* **3**, 285 (1936).
(14) Sörensen, C. *Ann. Phys., Lpz.*, **26**, 121 (1936).
(15) Claes, J., Errara, J., Sack, H. *C. R. Acad. Sci., Paris*, **202**, 1493 (1936).
(16) Baumgardt, E. *C. R. Acad. Sci., Paris*, **202**, 203 (1936).
(17) Sörensen, C. *Ann. Phys., Lpz.*, **27**, 70 (1936).
(18) Parthasarathy, S. *Proc. Ind. Acad. Sci.* **3**, 285, 482; 519 (1936).
(19) Hiedemann, E., Seifen, N., Schreuer, E. *Naturwissenschaften*, **24**, 681 (1936).
(20) Liepmann, H. W. *Helv. phys. Acta*, **9**, 507 (1936).
(21) May, J. Thesis, Durham (1937).
(22) Jeans, J. *Dynamical Theory of Gases*, p. 302 (1904).
(23) Herzfeld, K. F. and Rice, F. O. *Phys. Rev.* **31**, 691 (1928).
(24) Groszmann, E. *Ultraakustik*, p. 500 (1934).
(25) Henry, P. S. *Proc. Camb. phil. Soc.* **28**, 249 (1932).
(26) Belikov, P. *Z. phys.* **39**, 233 (1926).
(27) Richardson, E. G. *Proc. roy. Soc.* A, **146**, 56 (1934).
(28) Biquard, P. *C. R. Acad. Sci., Paris*, **202**, 117 (1936).
(29) Lucas, R. *C. R. Acad. Sci., Paris*, **203**, 459, 611 (1936).
(30) Mandelstam, L. and Leontovic, M. *C. R. Acad. Sci., U.R.S.S.* **3**, 111 (1936).

HEAT

By R. W. POWELL, B.Sc., Physics Department,
The National Physical Laboratory, Teddington, Middlesex

§ 1. PRODUCTION AND MEASUREMENT OF VARIOUS TEMPERATURES

The determination of temperature is undoubtedly the most important measurement in the subject of heat. The international temperature scale[1] which was agreed upon in 1927 was dealt with in the first volume of these *Reports*. Subsequent work by the standardizing laboratories has been directed towards the establishment of secondary points at the upper end of the scale, and the reproduction of the fixed points on the temperature scale with increased accuracy. Thus during the year now under review Schofield[2] has made a new determination of the freezing-point of palladium, obtaining a value of $1554{\cdot}4 \pm 1{\cdot}0°$ C., which is in good agreement with the value of 1555° C., originally recommended. Attention has also been devoted[3] at the National Physical Laboratory to that part of the scale from 660 to 1063° C. which is realized in practice by the standard thermocouple comprising platinum and platinum with 10 per cent of rhodium as calibrated at the freezing-points of antimony, silver and gold. The accuracy of calibration has been found to be limited by the variation of the e.m.f. with depth of immersion, due to temperature gradients in the ingot and inhomogeneity of the wires. Improved conditions are being obtained by designing a furnace in which the ingot is surrounded by a considerable mass of metal, and by preventing the thermocouple wires from being subjected to strain, by having them inclosed in a continuous sheath prior to annealing and during the calibration. A further innovation is the use of duplicate thermocouples joined in a common junction. By connecting the four leads in all six ways it is possible to appreciate when parasitic e.m.fs. are present.

The international temperature scale has not been defined below $-182{\cdot}97°$ C., the boiling-point of oxygen, and in view of the tremendous interest in low-temperature physics, it is not surprising that a considerable amount of work is being directed, both towards the establishment of lower fixed points, and the development of more convenient methods of temperature measurement than the standard gas thermometer.

Henning and Otto[4] have made a direct comparison between the platinum-resistance thermometer and the helium-gas thermometer over the temperature range 14 to 90° K. For this purpose they constructed a gas thermometer, within the measuring vessel of which was built a platinum-resistance thermometer, the latter being surrounded by a vapour-pressure thermometer. As a result of a careful series of measurements it was found that for the resistance thermometer the ratio of its resistance R_T at an absolute temperature T to its resistance R_0 at 0° C. could

be expressed by means of an empirical equation, which had been previously suggested by these workers,[5] viz.

$$10^5 R_T/R_0 = a + (bT + cT^2)\, D\,(100/T) + (d + eT - cT^2)\, D\,(230/T) \quad \ldots\ldots(1),$$

where a, b, c, d and e are constants for the particular thermometer, and $D\,(100/T)$ and $D\,(230/T)$ the Debye functions published by Beattie[6].

In the above-mentioned work measurements were made using hydrogen, neon, oxygen and nitrogen within the vapour-pressure thermometer also inclosed within the gas thermometer. From these results the eight fixed points given in table 1 are obtained[7].

Table 1. Fixed points in the range 14 to 90° K.

Gas	Normal boiling-point under 760 mm. of Hg		Triple point under p mm. of Hg		
	° C.	° K.	° C.	° K.	p
Hydrogen	−252·780	20·380	−259·20	13·96	54·1
Neon	−246·087	27·073	−248·60	24·56	325
Nitrogen	−195·808	77·352	−210·02	63·14	94·6
Oxygen	−182·970	90·190	−218·83	54·33	1·2

Five fixed points are required to enable the five constants of equation (1) to be determined. It will be seen that the eight points in table 1 are somewhat unevenly distributed over the temperature range. Henning and Otto[8] consider four of them to be suitable for fixing the equation of a platinum-resistance thermometer, namely, the normal boiling-point of oxygen (this forms a junction with the international temperature scale), the triple point of hydrogen, to define the lower end of the range, the normal boiling-point of hydrogen and the triple point of nitrogen. For the fifth point they suggest that use might well be made of the β-γ transformation of oxygen. Provided observations are made during heating they find a well-defined equilibrium temperature at 43·76° K., which can be maintained for about a quarter of an hour.

As the facilities of a specially equipped standardizing laboratory are necessary for determinations at these last two points, Henning and Otto[8] have directed their attention to the order of accuracy obtainable if two points only are used, such as the normal boiling-points of oxygen and hydrogen. They use this data to evaluate the equation

$$\Delta r_T = (T - 273)\,[A + B/(T + 10)] \quad \ldots\ldots(2),$$

where Δr_T denotes the difference between R_T/R_0 for the thermometer under test, and that for another resistance thermometer which has been fully calibrated in terms of equation (1), and A and B are the constants to be determined. Results are given for five pairs of resistance thermometers. In general the maximum error introduced by the adoption of this simplified calibration amounted to 0·14° within the range 20 to 90° K., but increased below 20° K. Furthermore, for a thermometer having an unusually high value for R_T/R_0 the error amounted to 0·7°. It is thus

apparent that the simplified method of calibration should only be used for thermometers constructed from similar wire to that of the comparison thermometer.

The difficulties introduced by the variations in the properties of different samples of platinum have also been discussed by Keesom and Bijl[9], who remind us that as long ago as 1914 Onnes and Holst suggested that gold would possibly prove more suitable than platinum in this range of temperature. Keesom and Bijl determined the {resistance, temperature} curves for four platinum thermometers over the range -190 to $-258°$ C. by comparing them with a helium-gas thermometer. They found equation (1) to be applicable to three of the thermometers, but the fourth, which had a low residual resistance, gave deviations amounting to $0{\cdot}15°$.

It might be observed that in a very recent publication Meissner[10] shows that the resistivity of a platinum wire at 20° K. depends, not only on its purity and the cold work introduced by the drawing, but also on the manner in which it is annealed. Minimum R_{20} values are obtained after annealing for about half an hour at a temperature of 750° C., whereas annealing at about 1300° C. gives values of the order of 30 per cent higher.

By reference to the two sets of temperature data given in table 1 it will be seen that the temperature of the ice point on the thermodynamic scale is accepted by the Reichsanstaldt workers as $273{\cdot}16°$ K. Keesom and Tuyn[11] have published a comprehensive account of work on gas thermometry which incidentally shows how recent determinations have become localized at this temperature. Since then Roebuck[12], from Joule-Thomson data for air and helium, has derived the value $273{\cdot}16 \pm 0{\cdot}02°$ K., whilst Kinoshita and Oishi[13], from observations of the expansion and pressure coefficients of gaseous nitrogen, hydrogen and helium, have obtained the value of $273{\cdot}16_7 \pm 0{\cdot}00_8°$ K. Jacyna[14], however, suggests the higher value of $273{\cdot}22°$ K.

The realization that water is not the simple substance it was thought to be when it was originally chosen to provide the first two fixed points on our scale of temperature, but has a complicated isotopic composition depending on its origin and treatment, does not introduce any very serious complication. This has been pointed out by Riesenfeld and Chang[15], who show that on the present scale of temperature the melting-point of pure light water $^1H_2{}^{16}O$ is $-0{\cdot}0014°$ C. and its boiling-point $99{\cdot}99945°$ C. Thus, if it were decided to define the degree in terms of pure $^1H_2{}^{16}O$ its value would be $1{\cdot}000008_5°$ C. The difference likely to arise between various samples of normal water is unimportant, being only $1{\cdot}0000001_3°$ C.; nevertheless, it would be preferable if in future definitions the standard substance were specified with greater precision.

With regard to the determination of the thermodynamic temperature scale at temperatures well below 1° K., where the usual forms of thermometry are no longer available, De Haas and Wiersma[16] find that this can be done from purely magnetic measurements on caesium-titanium alum. They have found experimentally that for this substance, not only does the intensity of magnetization σ remain constant when the magnetic field H is reduced from 24075 gauss at $1{\cdot}20°$ K., down to 104·6 gauss,

but also that this effect is independent of the initial temperature. For such a substance it is shown that

$$H = Tf(\sigma) + g(\sigma) \qquad \ldots\ldots(3).$$

When the saturation curves of the salt are known in the region of liquid helium, it is possible to find, on two of them, values of H and T such that the magnetization is equal to the magnetization σ_0 which remained constant during the demagnetization experiment. If these values are H_1, T_1 and H_2, T_2, then by equation (3) T_2 can be derived in terms of the other known values.

Those who desire information as to the equipment necessary for the production of liquid helium temperatures should refer to the extensive publications from the workers at Leiden. Rollin[17] has presented a paper to this Society describing a combined hydrogen and helium liquefier, which requires only liquid air for the initial cooling. Full constructional details are given for this apparatus, which embodies a small Linde liquefier for hydrogen, and an expansion method for liquefying helium, and is designed to enable low-temperature experiments to be made in the ordinary laboratory. By using about 9 l. of liquid air three complete expansions can be carried out, each producing about 1·6 cm³ of liquid helium, which lasts for between 1 and 2 hr., and is sufficient to cool normal-sized specimens. Kürti, Lainé, Rollin and Simon[18], by demagnetizing from a field of 32 kilogauss have been able to produce a temperature below 0·009° K. in ferric alum. Three of the last-mentioned authors have made some preliminary experiments[19] to investigate the time taken for the attainment of temperature equilibrium at low temperatures. They have shown that owing to the anomalous specific heats of the paramagnetic salts employed to furnish these temperatures, there is only a slow rate of temperature-exchange within the salt itself. They conclude from this that by increased subdivision of the salt it should be possible to cool helium to considerably below 0·1° K. within a reasonable time, unless the heat transmission should become controlled by the helium itself.

Turning our attention now to high temperatures at which the equilibrium temperature is mainly conditioned by radiant heat interchange, and, consequently, the emissivity of the hot surface is of prime importance, we find that more complete data have been determined for the emissivities of the metals used for high-temperature filaments. Ornstein[20] has published tables giving the emissivity of tungsten for wave-lengths of from 0·23 to 2·0 μ. over the temperature range 1600 to 3000° K., which show that there is a more than twofold variation in the emissivity for different wave-lengths. Bossart[21], working at one wave-length only of approximately 0·66 μ., has obtained emissivity values for tungsten and molybdenum, over the temperature range 1300 to 2000° K., which are in good agreement with Worthing's results, the maximum difference only amounting to 2 per cent. Although the emissivity of tungsten is only some 20 per cent greater than that of molybdenum, Bossart finds the emissivities of alloys of these two metals to be from 10 to 20 per cent in excess of the value calculated according to a simple mixing rule.

The high-temperature properties of niobium have been studied by Reimann and Grant[22]. They employed a very pure specimen, stated to contain only 0·02 per

cent of tin as impurity, and it is of interest to note that they give 2770° K. as the melting-point. This temperature was obtained from resistivity measurements made as the wire was heated to fusion, and is thus only an approximate value. It is 547° higher than that given in the *International Critical Tables*, from which it would seem that the melting-point of niobium is considerably affected by impurities. A precision determination on this metal will be awaited with interest.

Reference was made last year to the need of more convenient methods for the measurement of the temperatures involved in iron and steel manufacture. Naeser[23] has described a combined colour pyrometer and comparison lamp, which has been successfully used in Germany for measuring such temperatures. The light from the source, after passing through a wedge which only transmits complementary red and green colours, and an adjustable slightly tinted neutral wedge, produces a yellowish white field. This is matched against light reflected by a half-silvered mirror from a comparison lamp. The instrument is calibrated against a black body of known temperature. It is fitted with two scales, the colour temperature being given by the setting of the colour wedge, and the black temperature, in the red and green, by the neutral wedge. Both read the same on a black body. Sighted on a grey body the colour temperature is correct and the black about 50° C. low when a truly neutral wedge is used, whilst on a metal surface the colour temperature may be 20 to 40° above the true temperature and the black temperature 150° C. low. By slightly tinting the neutral wedge, as is done in the final instrument, it is possible so to adjust that on a grey surface the colour temperature is 10° low, whilst on a metal surface it is about 10° high. The reading on this scale thus gives the true temperature to within 10°, whilst the difference between the two scale readings enables some idea to be formed regarding the emissivity of the surface.

One advantage of optical methods for the measurement of surface temperatures is that the surface itself is not interfered with in any way. It appears that this advantage may shortly become available at temperatures below the present visible region. Improvements in the technique of infra-red photography have led Hencky and Neubert[24, 25] to apply this method to the measurement of temperatures over the range 350 to 750° C. The method requires a knowledge of the emissivity of the surface for the wave-lengths employed, but if it can be successfully developed it is one which should prove useful in certain instances, such as the determination of the temperatures of moving parts of a machine. For the purpose of studying the surface temperature of metals when rubbed one over another, Bowden and Ridler[26] have employed the ingenious method of forming the two metals into a thermo-couple. In this way they find that the observed temperature, which is that of the localized hot spots at which point contacts are made, rises to a maximum which corresponds to the melting-point of the more fusible of the two metals, and for high melting-point metals may exceed 1000° C. The work is furnishing useful information on polishing and lubrication processes.

Pfriem[27] has dealt with the practical measurement of rapidly varying temperatures. When the temperature to be measured is that of the cylinder wall of an engine he shows that the normal method of plugging the junction of a thermo-

couple into the wall becomes unsatisfactory when the temperature is rapidly changing. A better method is suggested in which a 2-mm. tapered plug of one element is inserted in the centre of a 5-mm. plug of a second element, but separated from it. The composite unit is fixed into a tapered hole in the cylinder wall so as to come some 0·05 mm. below the surface. A plate of a good conducting metal 0·05 mm. thick is soldered in this recess so as to make good contact with both couple elements, and when used in conjunction with a string galvanometer this device is considered to enable the temperature variations of the cylinder wall to be followed. The same author has made a theoretical study of the measurement of fluctuating gas and liquid temperatures by thermometers immersed therein[28]. An equation is obtained enabling the true {temperature, time} curve to be derived from those obtained experimentally using resistance thermometers composed of two wires of different diameter, heated by the same current, or using similar thermometers heated by different currents. He has applied the method to determine the variation with time of the temperature of the air within an air-compressor.

With regard to temperature control there have been the usual number of descriptions of thermostatic devices designed to suit special requirements, for details of which reference should be made to *Science Abstracts*. Of more general interest are two papers dealing with the theory of control systems. In the first, by Callender, Hartree and Porter[29], a study is made of the operation of control gear which is actuated by the property being controlled, and in which there is a time lag between the property-change and its response. The control equations are obtained and solved for specified conditions and values of the parameters involved. Based on these results, a practical method is outlined for obtaining automatic control of the type studied. The other treatise, by Turner,[30] arose out of an attempt to avoid hunting in a thermostatic control by making the slave-coil current a continuous function of the master-coil temperature. In practice, however, when the control sensitivity was sufficiently increased, violent thermal oscillation set in. A thermal model to represent this case has been devised, for which it has been found possible to calculate the critical value of the sensitivity, and to show that it is of the order obtained experimentally.

References

(1) *Trav. Bur. int. Poids Mes.* **18**, 94 (1927).
(2) Schofield, F. H. *Proc. roy. Soc.* A, **155**, 301 (1936).
(3) *Rep. nat. phys. Lab., Lond.*, p. 22 (1936).
(4) Henning, F. and Otto, J. *Phys. Z.* **37**, 601 (1936).
(5) Henning, F. and Otto, J. *Z. ges. Kälteindustr.* **39**, 86 (1932).
(6) Beattie, J. A. *J. Math. Phys.* **6**, 1 (1926).
(7) Henning, F. and Otto, J. *Phys. Z.* **37**, 633 (1936).
(8) Henning, F. and Otto, J. *Phys. Z.* **37**, 639 (1936).
(9) Keesom, W. H. and Bijl, A. *Physica*, **3**, 418 (1936).
(10) Meissner, W. *Ann. Phys., Lpz.*, **29**, 251 (1937).
(11) Keesom, W. H. and Tuyn, W. *Trav. Bur. int. Poids Mes.* **20** (1936).
(12) Roebuck, J. R. *Phys. Rev.* **50**, 370 (1936).
(13) Kinoshita, M. and Oishi, J. *Actes VII Cong. Int. Froid*, **1**, 155 (1937).

(14) Jacyna, W. *Z. Phys.* **97**, 107 (1935).
(15) Riesenfeld, E. H. and Chang, T. L. *Phys. Z.* **37**, 690 (1936).
(16) De Haas, W. J. and Wiersma, E. C. *Physica*, **3**, 491 (1936).
(17) Rollin, B. V. *Proc. phys. Soc.* **48**, 19 (1936).
(18) Kürti, N., Lainé, P., Rollin, B. V. and Simon, F. *C.R. Acad. Sci., Paris*, **202**, 1421 and 1576 (1936).
(19) Kürti, N., Rollin, B. V. and Simon, F. *Physica*, **3**, 266 (1936).
(20) Ornstein, L. S. *Physica*, **3**, 561 (1936).
(21) Bossart, P. N. *Physics*, **7**, 50 (1936).
(22) Reimann, A. L. and Grant, C. K. *Phil. Mag.* **22**, 34 (1936).
(23) Naeser, G. *Mitt. K.-Wilh.-Inst. Eisenforsch.* **18**, 21 (1936).
(24) Hencky, K. and Neubert, P. *Arch. tech. Messen*, **62**, T 100 (1936).
(25) Neubert, P. *Arch. tech. Messen*, **63**, T 114 (1936).
(26) Bowden, F. P. and Ridler, K. E. W. *Proc. Camb. phil. Soc.* **31**, 431 (1935); *Proc. roy. Soc.* A, **154**, 640 (1936).
(27) Pfriem, H. *ForschArb. IngWes.* **6**, 195 (1935).
(28) Pfriem, H. *ForschArb. IngWes.* **7**, 85 (1936).
(29) Callender, A., Hartree, D. R. and Porter, A. *Philos. Trans.* A, **235**, 415 (1936).
(30) Turner, L. B. *Proc. Camb. phil. Soc.* **32**, 663 (1936).

§ 2. CALORIMETRY AND SPECIFIC HEATS

Gas calorimetry. The subject of gas calorimetry, which was admirably dealt with a few years ago in a book by Hyde and Mills[31], received a further valuable contribution when Sir C. V. Boys chose as the title of the Guthrie lecture he delivered to this Society[32] "My recent progress in gas calorimetry". He considered the original Boys gas calorimeter to suffer from disadvantages in that the metal parts are attacked by the acid gases and the products of combustion, and that during use precautions have to be taken to maintain suitable temperatures of the circulating water and surrounding air. These have been remedied in the instrument now described, which should function for many years without needing repairs or replacements, and is now considered by its inventor to be suitable for precision determinations of the heats of combustion of gases. The corrosion problems have been solved by the use of a pyrex glass open burner, and glass chimney and combustion vessel, whilst the metal surfaces are protected on the gas side with a layer of bakelite varnish. To obtain uniformity of performance under all conditions a new type of water supply unit has been designed in which the necessary compensation is obtained by ensuring that the quantity of water used at any moment varies inversely as the volume occupied by the gas. The gas-supply unit is also of new design, mercury being used as the working fluid so as to avoid the saturating influence of the usual wet-meter. In the completed instrument, the temperatures are measured by means of differential electrical thermometers, which can be linked with a recording mechanism.

Keith[33] has described a recording gas calorimeter in which automatic compensation for atmospheric changes and disturbing heat losses is obtained in a totally different way. It consists essentially of two chambers and a heat exchanger. Air drawn through the instrument is heated in the first chamber by the test gas, and after passing through the exchanger, where its temperature is brought back to

the initial value, the air enters the second chamber where it is heated by a standard source of heat. The two chambers are fitted with differentially connected thermometers, and the heat supplied by the standard is varied to maintain a zero reading, the calorific value of the test gas can then be obtained in terms of the standard.

Calorimetry at low temperatures. Keesom and Miss Keesom[34] have measured the specific heat of solid helium and also determined its heat of melting. This is not the first determination to be made of the specific heat of solid helium, as Kaischew and Simon[35] derived a value for this two years previously, but the Leiden workers have made measurements to a lower temperature. The results of the two determinations compare reasonably well. A range of densities from about 0·20 to 0·22 g./cm.³ were covered, the specific heat being found to be larger for the smaller densities, and to increase with temperature more rapidly than Debye's formula would predict. The heat of melting increases from 0·835 cal./g. at 2·5° K. to 1·365 cal./g. at 3·5° K.

Reference was made in last year's *Report* to the explanation given by Prof. Mott for the high atomic heat possessed by nickel at low temperatures. The high paramagnetic susceptibility of platinum has led him to expect that for this metal the free electrons would again make an appreciable contribution to the heat capacity. Kok and Keesom[36] have now determined the atomic heats of platinum and copper from 1·2 to 20° K., and the results for platinum show this to be so. In neither case, in fact, is it possible to represent the atomic heat by a Debye function only. For copper the atomic heat is given by $C_v = D(335/T) + 0{\cdot}0001777T$, indicating that the heat capacity of the conducting electrons only slightly exceeds the value given by Sommerfeld's formula for free electrons; for platinum, up to 14° K., $C_v = D(233/T) + 0{\cdot}001607T$, showing, as confirmed by new measurements on nickel[37], that the heat capacity of the electrons is several times greater than that derived from Sommerfeld's formula for free electrons. Incidentally, Simon and Pickard[38] have made measurements on palladium, another transition metal, and find that for this the additional term is even greater, being $0{\cdot}0031T$ cal./g.

Debye's theory for the specific heat of solids agrees with experiments for most monatomic substances which crystallize with a cubic structure. Considerable departures are, however, shown by the alkali metals. The case of lithium has been studied by Simon and Swain[39], who found that, although the specific heat could not be represented by a simple Debye function, it was possible to express it as the sum of a Debye function having a characteristic temperature of 505°, and a Schottky function having a maximum at about 85° K. Pankow[40], however, from intensity-measurements of Debye-Scherrer patterns obtained with lithium derived a characteristic temperature of $352 \pm 12°$, which is in good agreement with 363°, as obtained by Grüneisen from electrical resistance measurements, and 354°, which Fuchs[41] derived from the values of the elastic constants. Pankow's measurements suggest a reason for the departure, since they show that a large proportion of the heat supplied to lithium is used up in increasing the atomic vibrations. The matter has been further discussed by Fuchs, who points out that Debye's theory neglects both the dependence of the velocity of the elastic waves on their direction within

the crystal, and the change of velocity with wave-length. The elastic constants of the alkali metals are shown to be much more dependent on the direction of measurement than are those of other cubic crystals, and, if dispersion is neglected, Fuchs shows that this anisotropy can account for the type of divergence from a Debye form observed for the alkalis. The influence of dispersion would be in the opposite direction, and Fuchs suggests that the success of Debye's theory, for such metals as copper and aluminium, may be due to the fact that in these metals there is a fair balance between the effects of dispersion and anisotropy, whereas, for the alkali metals, the effect of the anisotropy is preponderant. The foregoing instance provides additional support for the theoretical work of Blackman[42], which showed that, although the Debye theory has proved of such use in this subject, a perfect Debye curve is of rare occurrence and probably happens accidentally.

An anomaly of quite a different type has been found by Trapeznikowa and Schubnikow[43] to occur at low temperatures for several paramagnetic salts. The salts examined had been reported by Woltjer and Wiersma to undergo a sudden change in magnetic susceptibility. The specific heat curves resemble that of a ferromagnetic metal, in that at a particular temperature the heat capacity increases to a sharp maximum, and it would seem that both effects are connected with a spontaneous magnetization, and that the point at which this occurs is analogous to the Curie point. Anhydrous salts were studied and the effect was found to be greatest for ferrous chloride, and to occur in decreasing amounts in cobalt chloride, chromic chloride and nickel chloride. Agreement was not always obtained between the temperatures at which Woltjer and Wiersma found the susceptibility change to occur and that at which the peak occurred in the specific heat curve, a difference of 16° being noted with chromic chloride, whilst the nickel salt had two small peaks on the thermal curve, the position of the upper one depending on the heat treatment. Manganese chloride, although isomorphous with three of the foregoing salts, was found to behave normally.

Another low-temperature determination has been carried out by Bartholomé and Eucken[44], who have obtained values for the specific heats of hydrogen and deuterium in the solid and liquid states. The two solids give results in agreement with Debye's theory, the characteristic temperatures being 105° for hydrogen and 97° for deuterium.

Determinations at normal and higher temperatures. (*a*) *Metals.* Turning now to new determinations of the specific heats of metals from atmospheric temperature upwards, it is seen that Jaeger and his collaborators have applied their cooling by immersion method to obtain the specific heats of titanium[45], cerium and lanthanum[46] up to high temperatures. Each of these metals is considered to undergo one or more transformations, for which additional information is obtained from {resistance, temperature} curves and differential heat capacity curves. Ewert[47] has used a similar method to determine the specific heat of nickel over the range 0 to 1000° C., whilst results for nickel and silver have been given over the range 100 to 500° C. by Bronson, Hewson and Wilson[48]. These two metals, as well as quartz and β brass, were included by Moser[49] in what appears to have

been a very careful determination over the range 0 to 700° C. The method used by Moser differed in that the specific heat was derived from measurements of the heat supplied to the specimen and its rise of temperature when surrounded by an adiabatically controlled furnace. This method has advantages when dealing with substances which undergo transformation, on account of the greater certainty with which the form of the {specific heat, temperature} curve can be obtained in the neighbourhood of the transformation.

Earlier in the year Stoner[50] had attempted to calculate the specific heat of a ferromagnetic metal by treating this property as the sum of four terms involving lattice, electronic and magnetic contributions, and had concluded that further precise determinations of the specific heat of nickel were required before the success of the method could be assessed. It is, therefore, of interest to compare the recent data for nickel with the values Stoner derived, and considered, from his analysis of the results of Mme Lapp, Grew and Ahrens. This is done in table 2.

Table 2. Values for the specific heat of nickel (C_p), cal./g.

Observer	Temperature (°C.)									
	0	100	200	300	350	400	450	500	600	700
Lapp (1929)	0·1048	0·1125	0·1227	0·1386	0·1547	0·1272	0·1286	—	—	—
Grew (1934)	0·1057	0·1142	0·1241	0·1395	0·1557	0·1300	0·1300	—	—	—
Ahrens (1934)	0·1048	0·1133	0·1241	0·1378	0·1532	0·1325	—	—	—	—
Ewert (1936)	0·1011	0·1128	0·1252	0·1385	—	0·1267	0·1274	0·1281	0·1295	0·1310
Bronson, Hewson and Wilson (1936)	—	$0{\cdot}113_3$	$0{\cdot}123_7$	$0{\cdot}137_8$	0·150	$0{\cdot}127_0$	$0{\cdot}126_4$	$0{\cdot}126_1$	—	—
Moser (1936)	0·106	0·1115	0·1228	0·1357	0·149	0·1250	0·1246	0·1247	0·1258	0·128

Stoner's difficulty was that over most of the range the values of Grew were about 2 per cent higher than those of Lapp. This small difference in C_p leads to a much greater difference in the calculated magnetic contribution, and Ahrens' data tended to agree with the lower values, but were not conclusive, as they were relative to an assumed standard value at 20° C. It will be seen from the tabulated values that the new data have rather increased the choice of possible values. For most of the range Lapp's data are fairly well confirmed by those of Ewert and Bronson, Hewson and Wilson. Moser's data are, however, between 1 and 2 per cent lower than the results of the latter workers. The differences are actually small, and, depending as the values probably do on the composition and heat treatment of the specimen, closer agreement may not be obtainable. One does wonder, however, whether there is any significance in the fact that over most of the range Bronson, Hewson and Wilson's values for silver are also between 1 and 2 per cent greater than those obtained by Moser. It is of interest to observe that Mme Lapp[51] has now carried out measurements of a similar kind to determine the specific heat of iron up to 960° C. The peak of the specific-heat curve occurs at 757° C., the Curie temperature. Ashworth[52] has found that a sharp rise in specific heat, of the typically ferromagnetic kind, also occurs for manganese at a temperature of 350° C.

(*b*) *Gases.* Two determinations of the specific heats of gases using the sound-velocity method have been described. Hubbard and Hodge[53] worked at a constant temperature of 27° C. and varied the pressure of the gas from 1 to 60 atmospheres, whilst Sherratt and Griffiths[54] worked at atmospheric pressure and studied carbon dioxide over the temperature range 0 to 1000° C. In the one instance where these results can be compared the values obtained for γ, the ratio of the specific heat at constant pressure to that at constant volume, only differ by 4 parts in 1000. The results of Sherratt and Griffiths for the specific heat at constant volume are in good agreement with the *International Critical Tables* value for carbon dioxide at 0° C., but become about 15 per cent greater at temperatures of from 200° C. upwards. It is satisfactory to observe that these new values lie within 2 per cent of the value calculated for the specific heat from data derived from infra-red and Raman spectra. They thus furnish further experimental confirmation that the spectroscopic method yields results of sufficient accuracy for most practical purposes.

Evidence that the calorimetric measurement may not always be in complete agreement with the conclusions reached from spectroscopic and molecular data is, however, afforded by another recent determination. Rossini[55] has calculated the heat of hydrogenation of ethylene from earlier data which he had obtained for the heats of combustion of ethylene and of hydrogen and ethane. The value which he obtains, at a temperature of 25° C., is $32{\cdot}78 \pm 0{\cdot}13$ kcal./mol. This is in fairly good agreement with the value of $32{\cdot}64 \pm 0{\cdot}06$ obtained a year previously by Kistiakowsky, Romeyn, Ruhoff, Smith and Vaughan by means of a direct determination, and would suggest that von Wartenberg and Krause's 1930 value of $30{\cdot}6 \pm 0{\cdot}4$ is too low. The spectroscopic value obtained by Teller and Topley[56] is $31{\cdot}05 \pm 0{\cdot}3$, and Rossini suggests that part or all the discrepancy may be eliminated in their treatment, by suitable restriction in the ethane molecule of the rotation of the two methyl groups about the C—C bond, or alternatively, by replacing in the ethylene molecule the magnitude of the torsional frequency about the C—C double bond.

The calorie. The realization of the isotropic nature of water should also lead to the formulation of a more detailed definition of the calorie. Riesenfeld and Chang[15] state that the specific heat of $^1H_2{}^{16}O$ at 15° C. is 0·99999 of that of normal water. Since this reduction almost balances the increase in the size of the degree which would result if $^1H_2{}^{16}O$ were selected as the standard substance, it follows that values of the calorie defined in terms of either $^1H_2{}^{16}O$ or normal water would only differ by one and a half parts in a million.

References

(31) Hyde, C. G. and Mills, F. E. *Gas Calorimetry* (1932). London: Ernest Benn Ltd.
(32) Boys, C. V. *Proc. phys. Soc.* **48**, 881 (1936).
(33) Keith, R. W. *Mech. Engng, N.Y.*, **58**, 225 (1936).
(34) Keesom, W. H. and Keesom, Miss A. P. *Physica*, **3**, 105 (1936).
(35) Kaischew, R. and Simon, F. *Nature*, **133**, 460 (1934).
(36) Kok, J. A. and Keesom, W. H. *Physica*, **3**, 1035 (1936).
(37) Clusius, K. and Goldmann, J. *Z. phys. Chem.* **31** B, 256 (1936).
(38) Simon, F. and Pickard. *Properties of Metals and Alloys*, by Mott and Jones, p. 193 (1936).
(39) Simon, F. and Swain, R. C. *Z. phys. Chem.* **28** B, 189 (1935).

(40) PANKOW, G. *Helv. Phys. Acta*, **9**, 87 (1936).
(41) FUCHS, K. *Proc. roy. Soc.* A, **157**, 444 (1936).
(42) BLACKMAN, M. *Proc. roy. Soc.* A, **148**, 365 (1935); **149**, 117 (1935).
(43) TRAPEZNIKOWA, O. N. and SCHUBNIKOW, L. W. *Phys. Z. Sowjet.* Summer Number, June 1936, p. 6.
(44) BARTHOLOMÉ, E. and EUCKEN, A. *Z. Elektrochem.* **42**, 547 (1936).
(45) JAEGER, F. M., BOTTEMA, J. A. and ROSENBOHM, E. *Proc. K. Akad. Wet. Amst.* **39**, 912 (1936).
(46) JAEGER, F. M., ROSENBOHM, E. and FONTEYNE, R. *Proc. K. Akad. Wet. Amst.* **39**, 442 (1936).
(47) EWERT, M. *Proc. K. Akad. Wet. Amst.* **39**, 833 (1936).
(48) BRONSON, H. L., HEWSON, E. W. and WILSON, A. J. C. *Canad. J. Res.* **14**, 194 (1936).
(49) MOSER, H. *Phys. Z.* **37**, 737 (1936).
(50) STONER, E. C. *Phil. Mag.* **22**, 81 (1936).
(51) LAPP, Mme E. *Ann. Phys., Paris*, **6**, 826 (1936).
(52) ASHWORTH, J. R. *Proc. phys. Soc.* **48**, 456 (1936).
(53) HUBBARD, J. C. and HODGE, A. H. *Phys. Rev.* **49**, 194 (1936).
(54) SHERRATT, G. G. and GRIFFITHS, E. *Proc. roy. Soc.* A, **156**, 504 (1936).
(55) ROSSINI, F. D. *Bur. Stand. J. Res., Wash.*, **17**, 629 (1936).
(56) TELLER, E. and TOPLEY, B. *J. Chem. Soc.* p. 876 (1936).

§ 3. CONDUCTION

Conduction in liquids. This branch of heat conduction is one to which scant reference has been made in the earlier volumes of these *Reports*. In general, liquids are poor conductors and have conductivities of the same order as ebonite and similar solids classed as heat insulators. In many cases, however, the heat transfer through a liquid is considerably augmented by convection currents, and when the thermal conductivity of a liquid is being measured precautions have to be taken to ensure that no convection occurs. The experimental methods used are similar to those employed in the case of gases, a steady flow of heat being established through a thin film of liquid retained either between concentric cylinders or horizontal planes spaced a few millimetres apart.

Methods of the latter type have been used in the past year by Erk and Keller[57] and Bates[58] for the determination of the thermal conductivity of glycerine-water mixtures. The experimental arrangements adopted by these workers, however, differed appreciably. Erk and Keller used discs of fluid 11·7 cm. in diameter and 0·3 cm. thick, retained between horizontal copper plates held apart by small glass distance-pieces. The upper plate was electrically heated and the lower one water-cooled. In order to keep the lateral heat losses small the apparatus was set up within a Dewar vessel. The temperature drop through the liquid, which varied from 7 to 12° C., was measured by thermocouples attached to the copper plates. The heat flow was derived from the measured electrical input after applying a number of corrections amounting in all to about 5 per cent of the input. The chief uncertainty appears to be the heat loss through the Dewar vessel and the conduction in its glass walls. Bates retained the liquid in a bakelite vessel 38 cm. in diameter and surrounded this with heat-insulating material. The downward heat-flow through a central area 12·7 cm. in diameter was measured by means of a water-flow calorimeter, and as the outer annulus of liquid served as a guard-ring no corrections were

necessary. By taking these precautions to eliminate convection near the walls it was found possible to work with layers of water no less than 100 mm. deep, half this depth being used in the present work. As a fine-wire thermocouple stretched horizontally was used to measure temperatures within the liquid, conductivity values over the whole range from 10 to 80° C. could be obtained in one experiment. The exploration within the liquid indicated that there was no appreciable temperature discontinuity between the liquid and the hot and cold chromium-plated surfaces with which it was in contact.

The results obtained from the two series of experiments for water and pure glycerine are given in table 3 at temperatures of 10, 30 and 70° C. For purposes of comparison the values selected in 1929 for inclusion in the *International Critical Tables* are also shown, as well as other values published since these tables were prepared.

Table 3. Thermal conductivity of water and glycerine

Liquid	Investigators	Method	Thermal conductivity (cal./cm.-sec.-deg.) at		
			10° C.	30° C.	70° C.
Water	*I.C.T.* (1929)	—	$0{\cdot}00136_2$	$0{\cdot}00144_8$	0·00160
,,	Smith[59] (1930)	Cylinder	—	0·00144	0·00148
,,	Schmidt and Sellschropp[60] (1932)	Cylinder	$0{\cdot}00137_8$	0·00147	$0{\cdot}00158_8$
,,	Martin and Lang[61] (1933)	Plate (1 mm.)	$0{\cdot}00142_8$	0·00149	0·00162
,,	Nukiyama and Yoshizawa[62] (1934)	Plate	$0{\cdot}00138_8$	0·00144	0·00155
,,	Erk and Keller (1936)	Plate (3 mm.)	0·00139	0·00146	0·00160
,,	Bates (1936)	Plate (50 mm.)	0·00138	0·00145	0·00160
Glycerine	*I.C.T.* (1929)	—	$0{\cdot}00067_8$	0·00071	$0{\cdot}00078_8$
,,	Erk and Keller (1936)	Plate (3 mm.)	0·00064	0·00064	0·00064
,,	Bates (1936)	Plate (50 mm.)	0·00068	0·00068	0·00068

The two 1936 determinations on water are in good agreement, and lie within 2 per cent of the values which were accepted in 1929. Schmidt and Sellschropp's values, by the alternative cylindrical method, differ by less than this amount, while the Japanese data are in close agreement at the lower temperatures but become 3 per cent less at 70° C. The mean of the 1929 values and these four recent concordant sets of values gives thermal conductivities of 0·00138, 0·00145 and 0·00159 cal./cm.-sec.-deg. at temperatures of 10, 30 and 70° C. These should be very close to the true thermal conductivity. Incidentally, Schmidt and Sellschropp conducted their experiments up to a temperature of 270° C. and found the conductivity to remain fairly constant at a value of 0·00164 from 105 to 160° C., and then to decrease to 0·00143 at 269·2° C.

With regard to glycerine, although the two new determinations differ by some 6 per cent, both show the conductivity to be independent of the temperature. The large temperature coefficient previously attributed to glycerine is thus unconfirmed.

Klassen[63] has used an optical method to compare the thermal conductivities of two transparent liquids. The method consisted in establishing the same heat

flow through two superimposed liquid layers, and observing the angles, ϕ_1 and ϕ_2, through which a horizontal parallel beam of light is deviated on passing through the two liquids. A knowledge of the respective expansion coefficients, β_1 and β_2, and refractive indices, n_1 and n_2, is required, as thermal conductivity λ is given by $\lambda_1 = \alpha_1\beta_1\phi_2\lambda_2/\alpha_2\beta_2\phi_1$, where $\alpha = (n^2+2)(n^2-1)/6n$.

The cylindrical method has also been used in several recent experiments. Horák[64] determined the thermal conductivities of a number of oils at room temperature, whilst at the National Physical Laboratory[65] conductivities of calcium chloride brines have been studied. The conductivity was found to decrease as the concentration of the brine is increased, the value for a brine of density 1·38 g./ml. being 20 per cent lower than that of pure water.

Katharina Bartens[66] has determined the effective thermal conductivity of mixtures of solid spheres and oil. As the nature of the test necessitated using a considerable space between the cylindrical surfaces the effects of convection were eliminated by employing a range of temperature differences, and extrapolating to zero temperature difference. Even when metallic spheres of high conductivity were used, the effective conductivity of the mixture was only ten times that of the oil alone, and it was concluded that the thermal contacts between the spheres remained poor and varied with the nature of the solid surfaces.

A paper by Smith[67] contains results for the thermal conductivities up to 100° C. of several members of the paraffin and benzene series. In all cases the conductivity is shown to decrease with increase in temperature. At a particular temperature the conductivity of the paraffin series increases fairly regularly with increase in molecular weight, whilst the reverse happens with the benzene series. The curves in the latter case are less smooth owing, apparently, to the value for fluor-benzene being some 6 per cent high. The value for this liquid has been rechecked, but this may not remove the possibility of impurities affecting the determination. The paper also contains tables of existing thermal conductivity data for all liquids, and reviews the various theoretical and empirical equations which have been suggested to represent the thermal conductivity of liquids. Smith tested these equations by means of room-temperature results for twenty-eight liquids, and table 4 summarizes the results of his comparison. In these equations, λ is the thermal conductivity (cal./cm.-sec.-deg.), α the gas constant $2{\cdot}02 \times 10^{-16}$ erg/° C., v the velocity of sound in the liquid (cm./sec.), M the molecular weight, ρ the density (g./cm^3), C_p the specific heat at constant pressure (cal./g.-deg.), δ the mean distance between the edges of the molecules (cm.), c_1 the inner repelling force of molecules (g.), k the volume compressibility (cm^2/g.), Δ the mean distance of separation (cm.) of centres of molecules, assuming an arrangement which is cubical on the average, and ν the kinematic viscosity (centistokes).

A liquid which is interesting because of its unusually high thermal conductivity is helium II. The very high conductivity possessed by this form of liquid helium became apparent in the course of experimental work at both Leiden and Oxford. Keesom and Miss Keesom[68], when making a determination of the specific heat of liquid helium, concluded from the marked change which occurred in the form

of the heating curves below the lambda-point that the thermal conductivity of liquid helium must have suddenly increased. Rollin[69] in the course of the investigation referred to in § 1[18] found that the thermal insulation of the vessel containing the liquid helium became much impaired when the liquid helium was at a temperature below the lambda-point. As this did not occur when the liquid helium was in a sealed capsule, that is, when the possibility of the liquid helium II condensing on the walls of the connecting tubes of the apparatus was removed, it was attributed to the very high thermal conductivity of liquid helium II.

Keesom and Miss Keesom[70] have since published the results of a preliminary determination of the thermal conductivity of liquid helium II. The normal plate method had been used for liquid helium I, for which a value of 6×10^{-5} cal./cm.-sec.-deg. was obtained at 3·3° K., but proved unsuitable for helium II owing to its high conductivity. Eventually observations were made on the flow of heat along

Table 4. Results of comparison made by Smith of existing data at normal temperatures and equations for thermal conductivity of liquids

Author	Equation	Mean deviation %	Maximum deviation %
Theoretical			
Bridgman[71] (1923)	$\lambda = 2\alpha v\,(\rho/M)^{\frac{1}{3}}$	15·2	39
Kardos[72] (1934)	$\lambda = \rho C_p v\delta$:	23·4	104
	Using, as a first approximation, $\delta = 0{\cdot}95 \times 10^{-8}$		
	Using $\delta = c_1 k^{\frac{1}{3}}/\Delta$ (corrected by Smith)	12·5	
Empirical			
Weber[73] (1890)	$\lambda = 0{\cdot}00359 C_p \rho^3\,(\rho/M)^{\frac{1}{3}}$	20·1	51
	Replacing 0·00359 by 0·0043	14·8	41
Smith (1936)	$\lambda_{20} = 0{\cdot}000011 + (C_p - 0{\cdot}45)^3/155 + (\rho/M)^{\frac{1}{3}}/800 + v^{\frac{1}{3}}/10000$	6·7	25

a column of the liquid contained in a length of capillary tubing 0·6 mm. in diameter. At temperatures of 1·4 and 1·75° K. the conductivity of liquid helium II was found to be of the order of 190 cal./cm.-sec.-deg., which is some 200 times that of copper at normal temperatures, and no less than 3 million times as great as that of liquid helium I.

Satterly[74] has published a comprehensive survey of the physical properties of solid and liquid helium.

Solids at low temperatures. From the results of demagnetization experiments using a cylindrical rod of a paramagnetic salt Kürti, Rollin and Simon[19] have concluded that the thermal conductivity of chromium-potassium alum at 0·18° Curie is $2{\cdot}3 \times 10^{-5}$ cal./cm.-sec.-deg., whilst that of iron-ammonium alum decreases from $1{\cdot}8 \times 10^{-5}$ at 0·10° Curie to $0{\cdot}6 \times 10^{-5}$ at 0·07° Curie. These values agree in order of magnitude with extrapolations of de Haas and Biermasz's curves for quartz to these temperatures. For quartz de Haas and Biermasz[75] found that as the

temperature was lowered the conductivity measured in a direction parallel to the hexagonal axis increased to a maximum value of some 4·4 cal./cm.-sec.-deg. at 10° K., and subsequently became proportional to $T^{2\cdot5}$. Rock salt was found to behave in a similar manner, a maximum conductivity of about 1·3 cal./cm.-sec.-deg. being obtained at 7·5° K. and the conductivity at lower temperatures being proportional to $T^{1\cdot6}$. These two results are in marked disagreement with Peierls's[76] prediction that the heat conductivity of crystals should be inversely proportional to T^2.

The low-temperature thermal conductivities of several metals have also been studied by these workers, and by Bremmer and de Haas[77]. Lead, tin, mercury and indium all attain a maximum conductivity at a temperature a little below the supraconducting point, and with further decrease in temperature the conductivity decreases. For lead the maximum value is 0·55 cal./cm.-sec.-deg., which occurs at 7·9° K., and it is found for this metal, and also for mercury, that the thermal conductivity at a particular temperature, when the metal is in the supraconducting state, is less than when it has been brought back to its normal state, by application of a suitable magnetic field.

The thermal conductivity of copper at 14·5° K. is 14·3 cal./cm.-sec.-deg., and still appears to be increasing with decrease in temperature. A similar type of variation is found for tungsten, another metal which does not become supraconducting at the temperatures so far available. When the crystalline and electronic contributions to the thermal resistance of copper are separated in the manner suggested by Grüneisen and Goens[78] the crystalline conductivity is found to be proportional to $T^{-2\cdot4}$, which is in closer agreement with Peierls's T^{-2} relationship.

De Haas, Gerritsen and Capel[79] have measured the thermal conductivities of bismuth single crystals in the direction of the trigonal axis at temperatures in the neighbourhood of 20 and 80° K., and have studied the effect of an imposed magnetic field on the thermal conductivity. Since this can be assumed only to influence the electronic conduction, they have been able to determine experimentally the crystalline and electronic contributions (λ_g and λ_e respectively) to the heat conduction. For one crystal at 80° K., $\lambda_e = 0\cdot0079_4$ and $\lambda_g = 0\cdot0304$ cal./cm.-sec.-deg., whereas at 20·2° K., $\lambda_e = 0\cdot011$ and $\lambda_g = 0\cdot166$. Seeing that the crystalline portion exerts so great an influence it is not surprising that the high values of $1\cdot69 \times 10^{-8}$ and $5\cdot57 \times 10^{-8}$ are obtained for the Lorenz function at 80 and 20° K. respectively. For the Lorenz function as derived from the electronic thermal and electrical conductivities, however, values of $0\cdot35 \times 10^{-8}$ and $0\cdot39 \times 10^{-8}$ are obtained, whereas the theoretical value is $0\cdot58 \times 10^{-8}$.

Grüneisen and Gielessen[80] have made a similar study of the thermal and electrical conductivities of bismuth crystals, prepared in the form of cylindrical rods, for which the angle between the crystal axis and the axis of the cylinder varied from 2 to 86°. The results at a temperature of 79·6° K. for the specimen with only 2° between the axes are in reasonably good agreement with those obtained by the Leiden workers. For this specimen Grüneisen and Gielessen find $\lambda_e = 0\cdot0075$ and $\lambda_g = 0\cdot027$, whilst the Lorenz function obtained from the total conductivities is

$1{\cdot}6 \times 10^{-8}$. As the angle between the axes is increased these values become greater, but for an angle of 86° the electronic Lorenz function is still 23 per cent less than the theoretical value.

In general, it appears that the Lorenz function of the poorer conducting metals increases at low temperatures, whilst that of the good metallic conductors decreases.

Conductivities of metals at normal and high temperatures. Some of the effects which become more pronounced at low temperatures are still noticeable at atmospheric temperature; thus metals such as bismuth, antimony, German silver and Lipowitz alloy still possess high Lorenz functions which are decreasing with increase in temperature, whilst aluminium and nickel have somewhat low values which are increasing with increase in temperature.

During the past year Powell[81] has published the results of an investigation which shows that this convergence towards a more or less common value persists up to high temperatures. The six metals studied, which included monel metal, nichrome and four steels, had Lorenz functions at 0° C. ranging in value from $0{\cdot}76 \times 10^{-8}$ to $1{\cdot}18 \times 10^{-8}$. At 400° C. they ranged from $0{\cdot}65 \times 10^{-8}$ to $0{\cdot}81 \times 10^{-8}$, and at 800° C. were grouped between $0{\cdot}64_5 \times 10^{-8}$ and $0{\cdot}68 \times 10^{-8}$. In a more recent paper[82] the same author has given values for a sample of Acheson graphite. This material is seen to be quite a good thermal conductor, the value being about 40 per cent of that of copper. Its Lorenz function at 0° C. is more than two hundred times the value indicated by the electronic theory, but it is of interest to observe that the Lorenz function of graphite also decreases rapidly with increase in temperature, and at 800° C. is only some seventeen times as great as that of most metals.

A comparative method for the measurement of thermal and electrical conductivities of metals at room temperature has been described by Förster[83], which is claimed to be simple in application and rapid in attaining equilibrium. Similar-sized test and standard rods are mounted side by side in a vertical position with their lower ends in good contact with a block of copper. An exactly similar heating unit is attached to the upper end of each rod, and the energies required to establish the same temperature gradient in each rod are determined. Only a small temperature rise is used and adjustment is made by having a pair of differentially connected thermocouples located at corresponding positions near the top of each rod. The ratio of the conductivities of the two rods is then directly proportional to that of the respective energy supplies. The rods are unlagged and the surface losses are assumed to be equal. When comparing metals of fairly similar conductivity this assumption appears justified, as a cyclical comparison of magnesium against aluminium, zinc against magnesium and aluminium against zinc, gave a difference of only 1·8 per cent in the final value obtained for aluminium. Bollenrath and Bungardt[84] have employed an absolute longitudinal heat-flow method to measure the heat conductivities of some aluminium alloys used for pistons. The measurements cover the temperature range 50 to 350° C., over which an increase in conductivity is observed which differs from 10 to 18 per cent for the various rods tested.

A radial-flow method has been used by Krainer[85] to determine the thermal conductivity of a series of carbon steels up to about 600° C., whilst for measurements up to 1200° C. a method similar to that originally used by Honda and Simidu has been described by Maurer[86]. In this latter work Armco iron and a large number of chromium steels have been included, and it is seen from the results that whereas the conductivities at 30° C. range from a value 0·173 for Armco iron to 0·042 for the poorest conducting steel, at 800° C. all the values lie between 0·086 and 0·059 cal./cm.-sec.-deg. This convergence to a more or less common value is supported by other determinations, and a survey of all the available data for steels shows the conductivities at 800° C. all to lie between 0·054 and 0·086 cal./cm.-sec.-deg. Beyond this temperature, according to Maurer's results, the conductivity is either independent of temperature or has a small positive coefficient.

Another metal whose thermal conductivity has been the subject of recent investigation is tungsten. A knowledge of this property is required for calculations of the temperature of lamp filaments, where allowance has to be made for the cooling effects of the leads. Langmuir and Taylor[87] have made measurements on three thoriated tungsten filaments of differing lengths, and have derived values for the thermal conductivity at temperatures ranging from 240 to 600° K. They show that over this range the conductivity decreases from 0·414 to 0·314 cal./cm.-sec.-deg., and can be extrapolated to join on to the higher temperature results of Forsythe and Worthing. Michels and Martha Cox[88] have extended the theory to include the case of a filament attached to end-springs, and find the thermal conductivity of tungsten to decrease from 0·54 cal./cm.-sec.-deg. at 78° K. to 0·45 at 273° K. Baerwald[89] has also dealt with the theoretical determination of the temperature distribution along a heated filament.

Insulating and refractory materials. A number of workers have given new results for the thermal conductivities of slabs of insulating materials. Nukiyama[90] has published values from 10 to 60° C. for quartz, silica, bakelite, rubber, coal, slate, marble, granite and porcelain, which have been obtained by comparison with a slab of glass of known conductivity. Birch and Clark[91] have measured the thermal conductivities of several rocks and similar poor conductors up to 500° C. The specimens in the form of thin discs were sandwiched between metallic surfaces, and the correction for the imperfect contact between these surfaces and the test piece was obtained by extrapolation from measurements made in atmospheres of helium and hydrogen. Batch and Meissner[92] have used a plate apparatus in which the various surfaces were held in good contact by means of an external electro-magnet. By evacuating the apparatus instead of surrounding it with insulating material the time required for equilibrium to be attained has been reduced to about 20 min. when testing porcelain and 90 min. for a poorer conductor of hard rubber.

With regard to the heat insulation of buildings Beckett[93] has determined the effective air to air thermal conductance of tiled roofs from comparative experiments in which the heat loss through the roof of a small room was measured. The same author[94] has shown that the decrease in heat resistance of a wall-board due to absorption of water vapour from moist air does not normally exceed 5 per cent.

Only when persistently wetted by condensation or leakage is the insulating value seriously impaired.

A book devoted to the subject of heat insulating values of constructional materials has recently been published by Cammerer[95].

Holler[96] has adapted the radial heat-flow method for testing relatively good conducting refractory materials of the silicon carbide type up to temperatures of about 1200° C. The thermal conductivity of these materials was invariably found to decrease with increase in temperature, the rate of decrease being less at the higher temperatures. Over the temperature range 400 to 1200° C. the conductivity of the least porous and best conducting variety tested decreased from 0·088 to 0·045 cal./cm.-sec.-deg. It will be noticed that the conductivity at 400° C. is of the same order as that of cast iron at the corresponding temperature.

References

(57) Erk, S. and Keller, A. *Phys. Z.* **37**, 353 (1936).
(58) Bates, O. K. *Industr. Engng Chem.* **28**, 494 (1936).
(59) Smith, J. F. D. *Industr. Engng Chem.* **22**, 1246 (1930).
(60) Schmidt, E. and Sellschropp, W. *ForschArb. IngWes.* **3**, 277 (1932).
(61) Martin, L. H. and Lang, K. C. *Proc. phys. Soc.* **45**, 523 (1933).
(62) Nukiyama, S. and Yoshizawa, Y. *J. Soc. mech. Engrs, Japan*, **37**, 347 (1934).
(63) Klassen, T. *Techn. Phys. U.S.S.R.* **3**, 832 (1936).
(64) Horák. *Gesundheitsing*, **59**, 170 (1936).
(65) *Rep. nat. phys. Lab., Lond.*, p. 18 (1936).
(66) Bartens, Katharina. *ForschArb. IngWes.* **7**, 174 (1936).
(67) Smith, J. F. D. *Trans. Amer. Soc. mech. Engrs*, **58**, 719 (1936).
(68) Keesom, W. H. and Keesom, Miss A. P. *Physica*, **2**, 557 (1935).
(69) Rollin, B. V. *Actes du VII Congrès. Int. du Froid*, **1**, 187 (1937).
(70) Keesom, W. H. and Keesom, Miss A. P. *Physica*, **3**, 359 (1936).
(71) Bridgman, P. W. *Proc. Amer. Acad. Arts Sci.* **59**, 141 (1923).
(72) Kardos, A. *ForschArb. IngWes.* **5**, 14 (1934).
(73) Weber, H. F. *Ann. Phys., Lpz.*, **10**, 103, 304 and 472 (1880).
(74) Satterly, J. *Rev. Mod. Phys.* **8**, 347 (1936).
(75) De Haas, W. J. and Biermasz, T. *K. Onnes Lab. Leiden Comm. Suppl.* No. 82*b*.
(76) Peierls, R. *Ann. Phys., Lpz.*, **3**, 1055 (1929).
(77) Bremmer, H. and de Haas, W. J. *Physica*, **3**, 672, 687 and 692 (1936).
(78) Grüneisen, E. and Goens, E. *Z. Phys.* **44**, 615 (1927).
(79) De Haas, W. J., Gerritsen, A. N. and Capel, W. H. *Physica*, **3**, 1143 (1936).
(80) Grüneisen, E. and Gielessen, J. *Ann. Phys., Lpz.*, **26**, 449 (1936).
(81) Powell, R. W. *Proc. phys. Soc.* **48**, 381 (1936).
(82) Powell, R. W. *Proc. phys. Soc.* **49**, 419 (1937).
(83) Förster, F. *Z. Metallk.* **38**, 337 (1936).
(84) Bollenrath, F. and Bungardt, W. *Metallwirtschaft*, **15**, 368 (1936).
(85) Krainer, H. *Z. techn. Phys.* **17**, 281 (1936).
(86) Maurer, E. *Arch. Eisenhüttenw.* **10**, 145 (1936).
(87) Langmuir, I. and Taylor, J. B. *Phys. Rev.* **50**, 68 (1936).
(88) Michels, W. C. and Cox, Martha. *Physics*, **7**, 153 (1936).
(89) Baerwald, H. G. *Phil. Mag.* **21**, 641 (1936).
(90) Nukiyama, S. *Trans. Soc. mech. Engrs, Japan*, **2**, 344 (1936).
(91) Birch, F. and Clark, H. *Phys. Rev.* **49**, 479 (1936).
(92) Batch, H. and Meissner, A. *Z. techn. Phys.* **17**, 283 (1936).
(93) Beckett, H. E. *J. Instn Heat. Vent. Engrs*, **4**, 87 (1936).
(94) Beckett, H. E. *J. Instn Heat. Vent. Engrs*, **4**, 313 (1936).

(95) CAMMERER, J. S. *Die konstruktiven Grundlagen des Wärme und Kälteschutzes in Wohn- und Industriebau* (1936). Berlin: J. Springer.
(96) HOLLER, F. *Sprechsaal*, **69**, 733, 745 and 761 (1936).

§ 4. CONVECTION

Natural convection. It is usual when correlating the results of measurements of the heat transfer by natural convection to employ the following groups of dimensionless variables: Nusselt's number, $Nu = \alpha d/k$; Grashof's number, $Gr = gd^3\beta\theta/\nu^2$ and Prandtl's number $Pr = \nu/a$, where α is the heat transfer coefficient, in cal./cm.^{2}-sec.-deg., d is a length parameter, g the gravitational acceleration, θ the temperature excess, a the ratio of the thermal conductivity of the fluid to the specific heat of unit volume, and β, k and ν represent the coefficients of expansion, thermal conductivity and kinematic viscosity of the fluid. Hermann[97] directs attention to the need for a further dimensionless variable $Te = \theta/T_w$, the ratio of the temperature difference between the surface and the enclosure, to the absolute temperature of the enclosure, so that the complete relation is now represented by $Nu = F(Gr, Pr, Te)$. The influence of this new variable becomes appreciable in the region of small Grashof numbers of from 10^{-4} to 1, where, from an analysis of existing experimental data, Nu is shown to increase by about 22 per cent as Te is increased from 0 to 1. An accompanying theoretical analysis shows that an increase of about 19 per cent is to be expected.

Since $\nu = \eta/\rho$, where η is the viscosity of the fluid, it will be seen that Gr is proportional to ρ^2, and Saunders[98] has made use of this relationship to obtain information for the heat loss from a large surface by means of measurements carried out in compressed gases on relatively small surfaces. Results are given for four plane vertical surfaces having heights of 0·325, 2·5, 7·6 and 22·9 cm. immersed in air at pressures up to about 70 atmospheres. When plotted dimensionlessly the results for the four different heights do not deviate from the mean curve by more than 6 per cent. Most existing data agree well with the mean curve, although some give noticeably greater heat losses, possibly due to slight draughts or appreciable end-effects. It is of interest to observe that the heat transfer reaches a uniform value at a certain height from the base, after having passed through minimum and maximum values of approximately 0·65 and 1·1 times the final constant value. Griffiths and Davis[99] observed a variation of this nature some years ago, and considered that at first, where the air-motion was of a stream-line, or laminar, form, the warming of the air as it passed up the plate and the thickening of the laminar film caused the heat loss to diminish; in time, however, the development of turbulent air motion caused an increase in the heat loss which eventually reached a steady value. No maximum was then observed. Whereas Griffiths and Davis found the minimum to occur at a height within the range of Gr of from about $0{\cdot}15 \times 10^9$ to $1{\cdot}3 \times 10^9$, Saunders finds it to occur at about $2{\cdot}2 \times 10^9$, and considers that freedom from draughts accounted for the higher value. Hermann, in the paper to which reference has already been made, used the Schlieren method to observe the nature of the air flow over the surface of a vertical plane 1 m. in height. The

turbulence was observed to occur at a height corresponding to a mean value of about Gr of $1{\cdot}0 \times 10^9$. In these experiments the observed critical heights on the two sides of the plate differed appreciably, which would seem to indicate the presence of some such disturbance as that mentioned by Saunders. The individual critical values of Gr ranged from $0{\cdot}4 \times 10^9$ to $2{\cdot}0 \times 10^9$.

In connexion with this change in the nature of the air flow it is of interest to observe that Elenbaas[100] has attributed the unsteadiness which occurs in a high-pressure mercury-vapour discharge, when the pressure or the diameter exceeds a certain value, to the fact that the convectional flow then changes from a laminar to a turbulent form.

A review of the subject of heat transfer due to natural convection has been made by Missenard[101], who gives equations applicable to the region of laminar flow for the transfer from spherical surfaces and from vertical and inclined cylinders and planes.

Forced convection. When the hot surface is subjected to a fluid stream having velocity V, the fluid motions arising from local density differences become relatively unimportant, and Péclet's number $Pe = Vd/a$ replaces Gr, the heat transfer being given by $Nu = f(Pe, Pr)$. At times Reynolds' number, $Re = Vd/\nu$, is used instead of Pe.

A number of Russian workers have lately been studying problems of forced convection. Kroujiline[102] has obtained a theoretical solution for the transfer of heat from a plane surface subjected to a tangential stream of fluid, by introducing the idea of a thermal boundary layer through which the quantity of heat flowing varies inversely as the thickness of the layer. Approximations are made which hold for values of Re which are large, but not great enough for turbulence to appear. The resultant equation is $Nu = 0{\cdot}670 Pr^{\frac{1}{3}} Re^{\frac{1}{2}}$, which can be compared with $Nu = 0{\cdot}664 Pr^{\frac{1}{3}} Re^{\frac{1}{2}}$ previously derived by Pohlhausen, and $Nu = 0{\cdot}808 Pr^{\frac{1}{3}} Re^{\frac{1}{2}}$ obtained by Piercy and Schmidt. Incidentally, Piercy and Preston[103] have since employed a method of successive approximations to obtain a solution of the flat plate problem of skin friction and heat transfer at large Reynolds' numbers, whilst Howarth[104] has calculated the velocity and temperature distributions within the fluid.

Kroujiline[105] has since applied his treatment to the case of a cylinder set perpendicular to the stream, the final result being $Nu_\phi = A_\phi Pr^{\frac{1}{3}} Re^{\frac{1}{2}}$, where ϕ denotes the angle between the radius to the point considered and the direction of the stream and A_ϕ has appropriate values. Several workers have obtained results which show the total transfer from the cylindrical surface to be proportional to $Re^{\frac{1}{2}}$, whilst Kroujiline and Schwab[106] have obtained values of Nu_ϕ/Nu_0, which are in close accord with the present theory. Joukovsky and Kirejew[107] have investigated the temperature field around such a cylinder by a method of the Schlieren type in which the usual slit is replaced by a number of apertures in the form of a grating. The image formed under these conditions shows a series of bright and dark bands around the surface, whose distance apart depends on the temperature gradient at the point concerned. The calculated temperature distri-

bution from the optical field measurements is found to be in good agreement with actual determinations made with a thermocouple.

In many practical cases it is of importance to make the convectional heat transfer from a surface as great as possible. For some time past the addition of a series of suitably spaced radial fins to the surface of a pipe has been used as a means of increasing the heat interchange. More recently Büche and Schau[108] have investigated the advantages of having spiked projections on the surface, whilst Uhira[109] has advocated attaching a number of spiralled coils of wire to the tube surface. In the latter case a pipe 25 mm. in diameter, fitted with overlapping coils, which increased its overall diameter to 45 mm., was found to be about 40 per cent more efficient than a similar-sized pipe fitted with fins 45 mm. in diameter.

At other times it may be necessary to keep the heat interchange as low as possible, as when a body surface is protected by clothing. Angus[110] has carried out experiments which were designed to test the protection afforded by different types of normal covering materials, when either loosely or tightly fitted to the surface, for both still-air conditions and when exposed to a wind of about 5 miles per hour. Among the results obtained, it was found that for still-air conditions the normal indoor clothing of a woman is nearly as warm as that of a man, although the lighter covering affords much less protection against draughts.

Heat-transfer during fluid flow in pipes. This important aspect of heat-interchange between a fluid and the containing walls is one which in the past has formed the subject of many investigations. The subject is well surveyed in the new edition[111] of ten Bosch's well-known work on heat-transmission.

Jung[112] has employed hot gases flowing within a cooler pipe at velocities up to 420 m./sec. to study both the heat transfer and the frictional resistance. Brandt[113] has also shown how these quantities depend on the diameter and disposition of the pipe, whilst Shirokow[114] has extended existing theories to include the case of very high-velocity flow, where the fluid is appreciably heated as a result of the friction. He considers that in evaluating Nu it becomes necessary to use a fictitious temperature, equal to the sum of the normal mean temperature and the term $JV^2/2C_p$, where J is the mechanical equivalent of heat and C_p the specific heat at constant pressure. The equation connecting heat transfer with frictional resistance is found to be $Nu = \xi Pe/8$, where ξ is Nikuradse's frictional coefficient, and this has been confirmed experimentally.

References

(97) Hermann, R. *ForschArb. IngWes.* No. 379 (1936).
(98) Saunders, A. *Proc. roy. Soc.* A, **157**, 278 (1936).
(99) Griffiths, E. and Davis, A. H. *Spec. Rep. Food Invest. Bd, Lond.*, No. 9 (1922 .
(100) Elenbaas, W. *Physica*, **3**, 484 (1936).
(101) Missenard, A. *Chal. et Industr.* **17**, 3, 63 and 99 (1936).
(102) Kroujiline, G. *Techn. Phys. U.S.S.R.* **3**, 183 (1936).
(103) Piercy, N. A. V. and Preston, J. H. *Phil. Mag.* **21**, 995 (1936).
(104) Howarth, L. *Proc. roy. Soc.* A, **154**, 364 (1936).
(105) Kroujiline, G. *Techn. Phys. U.S.S.R.* **3**, 311 (1936).
(106) Kroujiline, G. and Schwab, B. *Techn. Phys. U.S.S.R.* **2**, 312 (1935).

(107) JOUKOVSKY, V. and KIREJEW, A. *Techn. Phys. U.S.S.R.* **3**, 754 (1936).
(108) BÜCHE, W. and SCHAU, F. *Arch. Wärmew.* **17**, 67 (1936).
(109) UHIRA, K. *Trans. Soc. mech. Engrs, Japan*, **2**, 373 (1936).
(110) ANGUS, T. C. *J. Text. Inst.* **26**, T 83 (1935) and **27**, T 273 (1936).
(111) TEN BOSCH, M. *Die Wärmeübertragung*, Auflage III (1936). Berlin: J. Springer.
(112) JUNG, I. *ForschArb. IngWes.* **7**, No. 380 (1936).
(113) BRANDT, H. *Wärme*, **59**, 367 (1936).
(114) SHIROKOW, M. *Techn. Phys. U.S.S.R.* **3**, 956 (1936).

§ 5. EVAPORATION AND CONDENSATION

Jakob, in the course of a series of lectures on evaporation and condensation, remarked that when he began to study the question of evaporation he was amazed at our general ignorance concerning it. The reading of the published account of these lectures[115] emphasizes the general truth of this remark, for they contain a wealth of information on these processes which has been made available by experiments carried out during recent years. The calorimetric experiments with which Jakob had been associated[116] for the determinations of the latent heat and specific volume of saturated steam are fully described.

Experiments relating to the boiling of liquids are then dealt with, and one learns that it was only six years ago that Bošnjaković[117] developed the first theory for the growth of steam bubbles within boiling water. He realized that this continuous one-way migration of molecules from the liquid to the vapour phase resulted from a slight superheating of the liquid which promoted the evaporation needed to form the bubble. Actually the amount of superheating remains very small provided there are sufficient nuclei present on which the bubbles can grow. Subsequent work by Jakob[118], Jakob and Linke[119] and Fritz and Ende[120], in which cinematographic records of the process were obtained and analysed, gave results for the heat transfer to the rising bubble which were in qualitative agreement with Bošnjaković's theory. The mean experimental value for this heat transfer coefficient ranged from 0·36 to 0·63 cal./cm²-sec.-deg., whereas Bošnjaković calculated the value of 1·01 cal./cm²-sec.-deg. At the moment of the bubble's origin the coefficient has the large value of 5·55 cal./cm²-sec.-deg. It is of interest to observe that the heat transfer to the rising bubble conforms to the laws of convection obtained from experiments on solids immersed in fluids. At first when the motion is stream line $Nu = 0{\cdot}56\ (Gr\ Pr)^{\frac{1}{4}}$, and when turbulence occurs $Nu = 0{\cdot}13\ (Gr\ Pr)^{\frac{1}{3}}$. Beyond $Gr\ Pr = 10^9$ higher values are obtained, owing, presumably, to the stirring effect of the bubbles, and then $Nu = 30\ (Hv/rw)^{0{\cdot}8}$, where H is the density of heat flow, v and r are the specific volume and latent heat of saturated vapour, and w is the rate of rise of the bubble.

Other large-scale researches into the thermal properties of steam have been in progress simultaneously at the National Bureau of Standards and the Masaryk-Akademie der Arbeit in Prag. The latest results of these investigations have been reported by Keyes, Smith and Gerry[121], and Havliček and Mišković[122], whilst new steam tables based on the results of this work have been published by Keenan and Keyes[123]. At the M.A.P. the total heat was measured and used to calculate the specific volume, whereas the American experiments were conducted in the

reverse way, the total heat being calculated from measurement of the specific volume. The same thermodynamic relation between these quantities was used in each case, and the results are in close agreement.

A number of workers, including Powell and Griffiths[124], Dörffel and Lettau[125], and Lurie and Michailoff[126], have studied the process of evaporation into a tangential stream of air from a water surface at a temperature well below the boiling-point. In the two former experiments information was obtained as to the variation of the rate of evaporation with length of surface (l). Both found the evaporation to be proportional to l^n, but whereas Powell and Griffiths gave $n=0{\cdot}77$, Dörffel and Lettau gave $n=0{\cdot}873$. Lurie and Michailoff subjected a water surface held in a pan 20 cm. in diameter to hot-air streams having temperatures ranging up to 225° C. Most other experiments have been made with heated water surfaces and air at atmospheric temperature. The rate of evaporation, however, is again found to be proportional to the difference in vapour pressure at the water surface and in the air at a distance. The data of Lurie and Michailoff agree fairly well with those of Hinchley and Himus, whose water surface was also bordered by the projecting edge of a pan, but are greater, by an approximately constant amount, than those of Powell and Griffiths, who endeavoured to avoid the turbulence set up by such an edge.

A type of wet condenser, or cooler, has been described by Zamenhof[127] in which the cooling pipes and fins are covered with a layer of porous material in communication with a tank of water. This method can be used as an alternative to the sprinkler system, and has the same advantage that as evaporation accounts for the greater part of the heat transfer, the temperature remains low.

Crout[128] has applied the kinetic theory to problems relating to the evaporation and sublimation of monatomic gases, and in collaboration with Giddings[129] has extended the treatment to gases having more than one atom per molecule. By evaporating aluminium within a vacuum and condensing the metal on a glass surface a good reflecting film is obtained. The development of the technique of this aluminizing process for the production of astronomical mirrors has been described by Strong[130].

References

(115) Jakob, M. *Mech. Engng, N.Y.*, **58**, 643 and 729 (1936).
(116) Jakob, M. and Fritz, W. *Phys. Z.* **36**, 651 (1935).
(117) Bošnjaković, F. *Tech. Mech. Thermo-Dynam., Berl.*, **1**, 358 (1930).
(118) Jakob, M. *Z. Ver. dtsch. Ing.* **76**, 1161 (1932).
(119) Jakob, M. and Linke, W. *Phys. Z.* **36**, 267 (1935).
(120) Fritz, W. and Ende, W. *Phys. Z.* **37**, 391 (1936).
(121) Keyes, F. G., Smith, L. B. and Gerry, H. T. *Proc. Amer. Acad. Arts Sci.* **70**, 319 (1936).
(122) Havlíček, J. and Miškovský, L. *Helv. Phys. Acta*, **9**, 161 (1936).
(123) Keenan, J. H. and Keyes, F. G. *Thermodynamic Properties of Steam, Including Data for the Liquid and Solid Phases* (1936). New York: J. Wiley and Sons.
(124) Powell, R. W. and Griffiths, E. *Actes VII Cong. Int. Froid*, **3**, 116 (1937).
(125) Dörffel, K. and Lettau, H. *Ann. Hydrogr., Berl.*, **64**, 342 and 504 (1936).

(126) Lurie, M. and Michailoff, N. *Industr. Engng Chem.* **28**, 345 (1936). See also *Gesundheitsing*, **59**, 289 (1936); **60**, 202 (1937).
(127) Zamenhof, S. *Rev. gén. Froid*, p. 92 (April 1936).
(128) Crout, P. D. *J. Math. Phys.* **15**, 1 (1936).
(129) Giddings, H. A. and Crout, P. D. *J. Math. Phys.* **15**, 124 (1936).
(130) Strong, J. *J. Astrophys.* **83**, 401 (1936).

§ 6. HYGROMETRY AND VAPOUR PRESSURES OF METALS

Hygrometry. The psychrometric chart published by Carrier[131], which has been widely used for the past 25 years in calculations involving the moisture content of air, was based on calculated values of adiabatic saturation, for it was believed that the temperature of adiabatic saturation was equal to that of a ventilated wet-bulb thermometer. Dropkin[132] has recently investigated the matter experimentally, and found that it is only at an air velocity of about 530 cm./sec. that this agreement occurs. At lower velocities the temperature of the wet bulb was above that of adiabatic saturation, and above 530 cm./sec. it assumed slightly lower values. This result was not unexpected, as it has been shown in a number of recent papers[133, 134, 135] that the Lewis relation, which led to an apparent agreement of the two temperatures, was based on a fallacy. It so happened that the relation was approximately true in the case of water, but fails completely for other fluids. Carrier and Mackey[136] have since made use of Dropkin's results, and of deductions from Arnold's psychrometric theory[137] and Carrier's own original measurements, in which the effect was observed but attributed to experimental error, to derive new tables for the psychrometric and thermal properties of moist air.

Two workers, Koch[138] and Powell[139], have studied the behaviour of fine-wire thermocouples when used for psychrometric purposes, and have shown that they possess advantages over the mercury-in-glass thermometers normally used, not only because they attain equilibrium more rapidly, but also because the full wet-bulb depression is obtained with very little forced ventilation. Koch, in fact, found that unventilated wet thermocouples reached the wet-bulb temperature, but those used by Powell indicated humidities about 2 per cent high under this condition. Simons[140], who used wet and dry thermocouples to measure very low humidities, found that a forced air velocity of only 2 cm./sec. was required to give the maximum depression. As the existing tables were considered inadequate for the low humidities dealt with in this investigation a comparison was first made between the readings of the thermocouples and the humidity, as determined from dew-point measurements. This involved a novel feature in that the dew was formed on a mercury mirror located on a rod of copper cooled at one end and fitted with a thermocouple to give the temperature of dew-deposition. Spencer-Smith and Matthew[141] used thermocouple psychrometers to test the performance of a small humatagraph of pocket size operated by the change in length of a short strip of fibre taken from the outer covering of the ovular bracts of conifers. They found this instrument to reach an equilibrium value in about 15 min., and noted that its behaviour became uncertain at 85 per cent humidity. The equilibrium water contents of such a fibre have been determined by Smith[142] at 15° C. throughout a cycle of humidity changes

from 10 to 98 per cent and down to 10 per cent again. In common with all materials of this type a hysteresis effect was observed, the difference in water content, at a particular humidity, depending on the direction of change. At 50 per cent humidity this difference amounted to 25 per cent for the cone fibre, compared with 17, 8 and 7 per cent for horsehair, raffia and goldbeaters' skin. Of these materials, the latter would appear to be the most suitable for use in this type of hygrometer.

Vapour pressures of metals. A critical survey of existing data for the vapour pressures of metals has been made by Eucken[143]. It is considered that some of the results obtained by boiling-point methods may have been seriously affected by the presence of absorbed gases from either the metal itself or the crucible, and that there is need for further determinations over the widest possible temperature range. The Neumann-Völker[144] modification of Knudsen's effusion method, in which there is no need to know the molecular weight of the vapour, is recommended for the measurement of low vapour pressures. In this method the substance under investigation is contained in a tube which is suspended by a quartz fibre. The vapour pressure can be determined from the angular deflexion which results when the vapour escapes through two suitably disposed small holes into an outer vacuum. For the measurement of high vapour pressures preference is given to condensation and entrainment methods. Although more exact data for the heat of fusion and the specific heat of the liquid metal are needed in many cases for deriving the vapour-pressure curve, $\log_{10} p = -A/T + B + C \log_{10} T + DT$, existing results have been analysed, and the constants A, B, C and D tabulated for each metal. The so-called Trouton's constant, the ratio of the molecular latent heat of vaporization to the boiling-point on the absolute scale at atmospheric pressure, is shown by these results to vary from a value of 16·1 for caesium to 29·3 for tin.

The vapour pressure of red phosphorus has been measured by Melville and Gray[145], who, employing an effusion method, obtained values about 10^7 times lower than those of the normal static method. They account for this by concluding that the effusion method allows no time for the combination of the molecules to form P_4, whereas in the closed system the equilibrium $P_4 \rightleftarrows 2P_2$ is set up. It was unfortunate that they considered the degree of extrapolation to be too great to give Trouton's constant with any degree of certainty, as the static method had indicated the abnormally high value of 37·4.

Potassium is another substance on which it appears desirable to have further information. Gordon[146] has pointed out that whereas the Neumann-Völker equation for the vapour pressure of this alkali metal leads to an entropy value in agreement with the third law of thermodynamics, the results of Fiock and Rodebush fail to do so, although he considers the latter to be better supported by spectroscopic data.

References

(131) Carrier, W. H. *Trans. Amer. Soc. mech. Engrs*, **33**, 1005 (1911).
(132) Dropkin, D. *Bull. Cornell Univ. engng Exp. Sta.* No. 23 (1936).
(133) Ackermann, G. *ForschArb. IngWes.* **5**, 95 (1934).
(134) Lewis, W. K. *Mech. Engng, N.Y.*, **55**, 567 (1935).
(135) Kirschbaum, E. *ForschArb. IngWes.* **7**, 109 (1936).
(136) Carrier, W. H. and Mackey, C. O. *Trans. Amer. Soc. mech. Engrs*, **59**, 33 (1937).
(137) Arnold, J. H. *Physics*, **4**, 255 and 334 (1933).
(138) Koch, W. *Gesundheitsing*, **59**, 504 (1936).
(139) Powell, R. W. *Proc. phys. Soc.* **48**, 406 (1936).
(140) Simons, A. *Proc. phys. Soc.* **48**, 135 (1936).
(141) Spencer-Smith, J. L. and Matthew, J. A. *J. Text. Inst.* **27**, T 219 (1936).
(142) Smith, A. J. M. *Rep. Food Invest. Bd, Lond.*, p. 219 (1935).
(143) Eucken, A. *Metalwirtschaft*, **15**, 27 (1936).
(144) Neumann, K. and Völker, E. *Z. phys. Chem.* **161** A, 33 (1932).
(145) Melville, H. W. and Gray, S. C. *Trans. Faraday Soc.* p. 1026 (1936).
(146) Gordon, A. R. *J. chem. Phys.* **4**, 100 (1936).

THERMODYNAMICS

By J. H. AWBERY, B.A., B.Sc., F.Inst.P.

§ 1. FUNDAMENTAL PRINCIPLES

The first law of thermodynamics. In the present article we propose to take stock of the present position of the science of thermodynamics, and it is therefore well to begin at the beginning, with the first law. Of its many forms, two may be selected for remarks here. In the first form, we express the law by saying "that whenever other forms of energy are converted into heat, there is a fixed ratio between the quantity of energy disappearing and the quantity of heat appearing". The truth of this statement has been so generally accepted that no definite attempts to prove or disprove it have been made for many decades, though efforts to measure the fixed ratio with greater accuracy have not been lacking, and the agreement between them (remembering that the mechanical energy dissipated has been of different types) is in itself evidence. The most recent redetermination[1] of the mechanical equivalent of heat was that carried out by Laby and Hercus in 1927, though it is understood that another determination is now in hand in Prof. Andrade's laboratory. Laby[2] gives a discussion of earlier determinations, from which table 1 is extracted.

Table 1. Mechanical equivalent of heat (hydrogen temperature scale)

	Erg per 20° cal.	Weight
Direct determinations		
Rowland	$4{\cdot}1822 \times 10^7$	2
Reynolds and Moorby	$4{\cdot}1758 \times 10^7$	1
Indirect electrical determinations		
Griffiths	$4{\cdot}1904 \times 10^7$	1
Schuster and Gannon	$4{\cdot}1898 \times 10^7$	1
Callendar and Barnes	$4{\cdot}1795 \times 10^7$	2
Bousfields	$4{\cdot}1767 \times 10^7$	1
Jaeger and Steinwehr	$4{\cdot}1821 \times 10^7$	2
Weighted mean	$4{\cdot}182 \times 10^7$	

Their own value for the 20° C. cal. is $4{\cdot}1809 \times 10^7$ ergs, corresponding to the value $4{\cdot}1860 \times 10^7$ ergs for the 15° cal.

The second form in which we shall state the first law of thermodynamics is: "A quantity, the internal energy, can be defined for any system and is a function of the state of the system only; the increase of the internal energy in passing from state A to state B is the sum of the heat absorbed by the system and the work done on it (measured in the same units) in passing from the one state to the other." We may note that since the internal energy U is a function of state only, the change ΔU referred to is independent of the path which leads from A to B, and moreover, there is no need for the path to be a reversible one. The last point is one which is not always clearly realized.

The first law is in some sense a generalization of the law of conservation of energy, which was originally formulated for conservative mechanical systems, and then extended to include electromagnetic forces. With Einstein's relativity theory of 1905, the law took on a new complexion, since it then appeared that energy has mass, so that in a sense, the conservation of energy became the conservation of mass. Normally the amount of energy concerned in our problems is too small to have appreciable mass, but in nuclear reactions the energy interchanges are so large that they are comparable with the masses of the reacting particles. The evidence, particularly in the case of the elements which emit β particles spontaneously, seems to point to some error in the foundations of the theory, since it seems proved that energy is not conserved in the individual disintegrations[3]. This might be due to a failure of Einstein's law, or to the existence of a form of energy of which no account has been taken (including the possibility that two atoms in different states may appear to us to be in the same state), or finally to a failure of the conservation law. If the latter is the explanation it throws doubt on the absolute accuracy of the first law itself, though the doubt is not serious, since there is every reason to suppose that even in the nuclear processes the law of conservation is true statistically even if not in individual acts. If either of the other explanations is the true one, it leaves the law of conservation true for the individual processes, and therefore raises no difficulties as far as pure thermodynamics is concerned. Since the first law is closely related to thermochemistry, we may here give references[4] to reviews or contributions to that subject by Rossini and by Roth.

The second law of thermodynamics. Like the first law, the second is capable of expression in many forms, of which this time we shall select three. The first has many wordings, the common feature of which is that they point out the fact that heat always runs "downhill", and cannot be "raised" without either the expenditure of work or the "lowering" of some other compensating quantity of heat. This is the law which expresses the generally accepted impossibility of using, for example, the heat stored in the sea as a nearly permanent source of energy. It does not forbid us to conceive of a machine taking heat from the equatorial seas and rejecting heat at the poles, and indeed we know by experience that such an engine works, for we see it daily in action in the atmosphere. Again, it does not forbid us to hope that the solar energy may in time be used directly, instead of being first stored by chemical action in vegetation. The refrigerating plant is no exception to the rule, since it either takes in heat, which is lowered to the temperature of the condenser, or has work done on it.

Of the other forms of the law, one group fixes attention on a machine (or a complete process), and the other focuses on the material undergoing the process. From the first point of view, we have the theorem as to the efficiency of the perfect engine, which leads immediately to the definition of the absolute scale of temperature. In this connexion, it may be worth noting that a Committee now sitting has considered the experimental evidence available, and proposes[5] the value 273·15 for the position of the ice point on the absolute scale, or Kelvin scale as it is generally called.

When the system, rather than the process, is placed in the foreground, we obtain from the second law of thermodynamics the statement: "A quantity, the entropy, can be defined for any system and is a function of the state of the system only; the increase of the entropy in passing from state A to state B is the sum of the elementary quantities of heat absorbed by the system in the passage, each divided by the absolute temperature at which it is received." Since the entropy S is a function of state only, it would appear that the changes ΔS in the entropy referred to should be independent of the path which leads from A to B, but actually this is only true for reversible paths, since the temperature at which heat is received is indefinite unless the source of heat is at sensibly the same temperature as the system when the transfer occurs. It is in this sense that the calculation of entropy changes is restricted to reversible paths.

Thermodynamic functions. We have seen that the first law defines U, the internal energy, and the second defines S, the entropy. With these it is customary to associate three other quantities which are functions of state only; the enthalpy $H = U + pv$, where p is the pressure and v the volume, Helmholtz' free energy $F = U - TS$, where T is absolute temperature, and a quantity now generally called Gibbs's function, $G = U - TS + pv = F + pv = I - TS$.

All these quantities except entropy have alternative names, which are still commonly found in the literature, and those given above were all permitted in 1934 by the International Union of Pure and Applied Physics. Nevertheless, the others are so common that the reader of thermodynamic literature must be familiar with at least the following:*

Internal energy is also called intrinsic energy, and is frequently (especially by engineers) denoted by E.

Enthalpy is also called total heat, and heat content, and is frequently denoted by I.

Helmholtz' free energy is often referred to without the qualification "Helmholtz" and is also called the maximum work or thermodynamic potential at constant volume. Many different symbols have been used for it, and in particular Lewis and his school (as well as Nernst†) use A.

Gibbs's function. This name is fairly recent. It used to be most often called thermodynamic potential at constant pressure, or even simply the thermodynamic potential, but in the early twenties of this century the most important American school of physical chemists christened it the free energy and gave it the symbol F, claiming that it had frequently been used even when the other quantity was intended. This procedure led to confusion among those not of the school mentioned, and led them to assign the name of Helmholtz to *their* free energy, whilst the name "Gibbs's function" presently appeared and might have offered a happy solution.

* A joint committee of the Chemical, Faraday and Physical Societies, after due consideration, now proposes to give up F for the Helmholtz free energy and to use A for this quantity and G for Gibbs's function, it being understood that the Americans, as a *quid pro quo*, will likewise renounce the symbol F and adopt G for their free energy.

† It is to be noted, however, that Nernst frequently uses A to refer, not to the integral quantity, but to the *change* in it which occurs during a chemical action.

It has been widely adopted, but it is well to state clearly that in the papers of G. N. Lewis and his pupils, the term—free energy—and the symbol F still denote Gibbs's function.

This multiplicity of functions is not due to reckless breeding: each has its part to play, and remains constant in some important group of processes. Thus the enthalpy is constant during the unrestricted expansion of a gas, its temperature-derivative for a simple substance is the specific heat at constant pressure, whilst the heat of reaction in a chemical process occurring at constant pressure (e.g. in an open vessel) is equal to the difference between the enthalpy of the products and that of the reactants. Gibbs's function has for the engineer the important property that its numerical value per unit mass is the same for a liquid and its vapour at the same temperature, since it is unaltered by an isothermal isobaric process. Consequently, it is unchanged in a chemical reaction occurring under these conditions.

If we consider unit mass of a substance, it is possible to imagine tables drawn up showing each of these functions when the substance is in all possible states. There is, however, a difficulty in that there is no definitely fixed state in which any of the functions is zero; in other words, we can only know *differences* in the functions. For the internal energy U, it seems reasonable to take the standard state at absolute zero of temperature, since we cannot expect to remove any more energy by cooling the body. But even this choice is not definitely dictated, and we can imagine that there may be stores of internal energy in the body even at this temperature. If Pauli's exclusion principle is remembered, for example, we see that all the electrons in an atom cannot be in the lowest orbit, and hence some must possess energy. The like may be true more generally, and is shown to be so in the Fermi-Dirac statistics. In any case, this does not deal with the question as to the pressure at which the internal energy is to be taken as zero. Presumably it would have to be zero pressure, since otherwise work could be obtained by expansion. If this choice is indeed made, then all the functions except possibly S vanish at the standard state, since the term U vanishes whilst pv vanishes because $p=0$ (unless v tends to infinity in such a way as to make pv finite) and TS because T is zero.

Actually, for engineering purposes, another choice is made, viz. the standard state is that in which a fluid is at 0° C. and under the saturation pressure of its vapour at this temperature.

The activity function. There are many quantities which "ought" to depend on concentration, but fail to do so exactly except when the latter is very small. Thus, the equilibrium constant K_c of the chemical reaction $A+B=C+D$, in which no solids take part, is $[C]\,[D]/[A]\,[B]$, where brackets represent concentrations. According to the law of mass action, K_c should be a constant at any given temperature, whereas experiment often reveals cases where it is not so. If, however, we examine the reasoning by which the mass-action law purports to be established, we find either that it is based on a statistical mechanics in which interactions between molecules and the volumes of the latter are ignored, or else it proceeds by expressing the fact that Gibbs's function is unchanged as the reaction proceeds[6]. In the latter case, the final result rests on assuming an expression for Gibbs's

function which equally fails to take account of molecular interaction and of the finite volume of molecules. Indeed, in the case of solutions, there are also electrical attractions and repulsions which are neglected.

The exact allowance for these effects is really beyond present theory, and hence the most convenient method of proceeding is to restore the accuracy of such deductions as Raoult's law and the law of mass action by replacing the concentration by another quantity, a sort of effective concentration. Among the various possible choices, the most convenient, and the one made by G. N. Lewis in 1907 when he introduced this notion[7], is that which makes the activity a kind of generalized concentration (in solution) or pressure (in a gas). Thus for a gas, the activity a is defined by $p_2 - p_1 = RT \log (a_2/a_1)$. This only serves to define differences, since it makes $p = (RT \log a) + b$. To complete the definition, the arbitrary additive constant b is defined by the condition $a \to p$ as $p \to 0$.

In the case of concentration cells, the electromotive force is found to be given incorrectly by the formulae which seem to stand on a sound theoretical basis, if the ionic concentrations are deduced from the conductivity in the usual manner. Here again an activity is used instead of concentration, so chosen that it approaches the concentration in sufficiently dilute solutions. Then E, the e.m.f., is accurately given by $(RT/nF) \log (a_1/a_2)$, where R and T have their usual meanings, whilst n and F are respectively the valency of the ion, and the quantity of charge known as 1 faraday, a_1 and a_2 of course being the activities of the concentrated and dilute solutions respectively.

In the case of solutions, it will be realized that this definition of activity is equivalent to defining the activity as an effective concentration such that the osmotic pressure is given accurately as a function of this "concentration" by the law of the perfect gas. Consequently, activities in solution can be deduced from freezing-points, vapour pressures, osmotic pressures, chemical equilibria or e.m.f. measurements, and much modern work on solutions can thus be conveniently correlated in this way.

Finally, for convenience of reference, we may mention that the *activity coefficient* frequently encountered in the literature is the ratio of the activity to the concentration itself, and therefore tends to unity in dilute solution.

The statistical aspect of thermodynamics. The state of the energy which is stored within a system—let us now restrict ourselves to a homogeneous system—is not enquired into by classical thermodynamics, but we know that it must in fact be possessed by the molecules, either as kinetic energy or as some form of potential energy. To study the matter statistically it is necessary to make more definite assumptions as to molecular constitution, but in return we obtain more definite information than can be obtained by classical methods.

On the other hand, the assumptions are by no means as definite as those made in purely molecular theories, and most results in statistical mechanics would remain true for whole groups of possible laws of interaction. It is, of course, necessary to make assumptions (which ultimately can only be judged by their results), and to explain them it will be necessary to remind the reader of the

methods of statistical mechanics. If a system, say a gas, in which molecules have no internal energy, has all its molecules in assigned positions, and moving in assigned directions with assigned velocities, we say that it presents a certain *complexion*; any alteration of the details will give another complexion, and in particular, if one molecule exchanges position and vector velocity with another, a new complexion will be presented. On the other hand, this exchange alters nothing *observable* in the gas, and the *statistical state* is therefore unaltered; if an assigned number (no matter which individuals) of molecules are in assigned positions and moving with assigned velocities, we have a particular *statistical state*. The fundamental assumption, then, is that the statistical state which occurs most frequently on enumerating all possible combinations of complexions is also most probable from a physical point of view. Indeed, this can be proved[8], if we regard all complexions as equally probable (or rather if the assumption corresponding to this is made, when we change from speaking of velocities and position to ranges of velocity and regions of space). This last change is frequently made by the introduction of a six-dimensional phase space, in which the six coordinates are the components of position and of velocity of the molecule. The postulate is then made that the point representing the condition of any molecule is as likely to be in one cell as in another, if the whole diagram is divided into cells of equal six-dimensional volume. In quantum mechanics a special division of the phase space into cells has to be made, and the differences between the Bose-Einstein, classical, and Fermi-Dirac statistics are that in them the presence of a representative point in a cell respectively increases, does not affect, and decreases, the likelihood of a second point being found there[9].

When this method of investigation is adopted, it is found that the statistical state of greatest probability is overwhelmingly so under ordinary conditions. In other words, we find that the equilibrium state is a perfectly definite one. For a gas in which no molecular forces are active, the space distribution is that of Maxwell's law. In fact, the number of molecules with energy w_r is $Ae^{-\mu w_r}$, where A is a constant, and μ is $3N/2E$, N being the total number of molecules and E the total energy (the internal energy). Further, if a mixture of two such gases is considered, the theory at once shows that they have the same μ. It is thus strongly suggested that this is closely related to the temperature as introduced into classical thermodynamics, for it is the same in a mixture of two gases, and is a function of the average energy of the molecules. Further confirmation is obtained by calculating the pressure of a hypothetical perfect gas, and the evidence is in fact conclusive[10] that μ is kT^{-1}. The next step in the theory as usually presented is to study the case where the molecules have energy of mutual position due to intra-molecular forces, and where they have internal energy. For the theory to be justified, it must, of course, be possible for the energy of one form to be changed into another, but there is no need to postulate any particular mechanism, provided only that some means of effecting the exchange is available. The simplest example will show that in the absence of this possibility there would be no justification for thinking that the statistical state obtainable most frequently from the complexions was the most

probable, for it might be unattainable. Thus a gas in which the molecules moved vertically between two walls so smooth that they reflected the impinging molecules back along the same paths could never reach the Maxwell distribution, and would exert no pressure on the side walls. When the necessary (and very plausible) assumption is made that these interchanges of energy are possible, the theory gives at once the extremely important law of the equipartition of energy. It also brings in again the ubiquitous μ, in the form of the Boltzmann factors, according to which the number of particles having energy of any type w_r is, in equilibrium, given by the same formula as we found for energy of translation, viz. $Ae^{-\mu w_r}$.

The statistical interpretation of entropy. The theory can be attacked from a slightly different angle, which proves to be somewhat superior, especially in more complex systems. Instead of counting complexions, let us imagine a whole ensemble of different gases, any one of which might be the one under consideration, i.e. all having the same volume and energy, but differing in their complexions. So far, this concept is identical with the first. But now, we imagine a much more immense phase space, viz. one in which a single point represents the whole of any one gas, the dimensionality of the space being so high ($6n$ if there are n molecules each with 6 degrees of freedom) that a single point represents all the coordinates of all the molecules of that gas both as regards velocity and position. Thus, as soon as a collision occurs in the gas, the complexion is altered and the representative point for that gas moves; our ensemble of gases is thus represented by a swarm of representative points, each moving on its own trajectory. Then it is a matter of straightforward dynamics to prove that if the points are originally distributed as a definite function of the energy, their density does not change with time provided the coordinates chosen are those of momentum and not of velocity[11] This shows that the points do not tend to crowd into any particular region, and thus gives some justification for supposing that any one system selected at random from the ensemble will be typical of the whole, and thus that any features common to the ensemble may be expected to reflect permanent properties of a real gas.

This method, too, leads at once to distributions described by the Boltzmann factors, and enables statistical mechanics to deduce the existence of a quantity which is a function of state only and which for a perfect gas varies with pressure and with temperature in the way which pure thermodynamics predicts for the entropy. Moreover, this quantity is found to increase in any spontaneous process. It is thus identified with entropy. Statistical mechanics, however, is here of greater power than classical thermodynamics, since it meets no difficulties due to irreversible changes. It also suggests, without establishing, a means of determining absolute instead of relative entropies.

The simplest, and far from rigorous, treatment of the matter was given by Boltzmann[12], the argument being that since any spontaneous change must be towards a state of higher probability, and since entropy increases in such a change, therefore probability and entropy increase or decrease together. Consequently he assumes that entropy must be a function of probability, and since probabilities in successive changes will be multiplied together, whereas the entropy changes will

be additive, it follows that the entropy S is of the form $A+B\log W$, where W is the probability and A and B are constants. This conclusion is generally accepted, and the absolute entropy is assumed to be obtained by taking $A=0$. The weakest point comes later, when the quantity W is identified with the number of complexions associated with a given statistical state. Foremost among the objections is the one that this is not a probability at all. It is the numerator in an expression which might equal the probability if the denominator were known. The difficulty is less acute since the advent of the quantum theory than it was previously, since in quantum theory, the poly-dimensional phase space is divided into cells of finite volume; all points in a single cell are regarded as equivalent, and all we need do is to take W as the total number of complexions of the required energy. The finite size of the phase cells prevents this number from being infinite, as it was in pre-quantum theory. Next, we have the difficulty, admittedly common to the whole of statistical mechanics in the first of the two forms described, that we have no guarantee that probability in this sense—the frequency sense—is necessarily identical with probability, in the sense of likelihood of finding that a physical system is in the condition concerned.

The most famous attempt to meet this difficulty is the assumption—called by Boltzmann the ergodic hypothesis, and by Maxwell the principle of the continuity of path—that if left long enough a system will pass through all the states possible to it[13]. If this is accepted, then we can be sure that for an overwhelming part of the time the system will be in or exceedingly close to those states which give its macroscopic properties their normal values—e.g. the molecules will be approximately uniformly distributed and their velocities will follow Maxwell's law. Unfortunately, we have no *a priori* reason to believe in the principle, except an instinctive feeling that the numerous collisions are bound to lead to all sorts of configurations and so to all the various possible states. The difficulty has intrigued many acute minds, and the discussion of alternative, perhaps more plausible, assumptions, given by Tolman is well worthy of perusal.

A defect of the usual statistical mechanics, of an entirely different nature, has been pointed out by Darwin and Fowler[14], and is so serious that it has led them to alter the whole presentation—though not any of the results—of the theory. It is in connexion with the use of Stirling's theorem for approximating to the factorial expressions which occur. The difficulty only arises when quantum restrictions are acknowledged, and is due to the fact that instead of our having a large number of molecules to partition among the different cells in the phase space, as is usually assumed, we have actually far more phase cells than molecules; in the case of say 20 l. of gas at 20° C. and atmospheric pressure, there are about 10^{24} molecules and 10^{63} cells in the position-momentum space, if each is of the volume h^3, where h is Planck's constant. Thus the number of complexions with n_1 molecules in cell 1, n_2 in cell 2, and n_c in the last of the C cells, is certainly $n!/n_1!\,n_2!\ldots n_c!$, but it is evident that we cannot be justified in expanding every one of these factorials. Some at least of the n_r must be too small (in fact zero) since Σn_r is $n=10^{24}$, whereas C is 10^{63}. Darwin and Fowler therefore use an entirely different method of

approximating to the number of complexions, which is in effect a calculation of average values in the system, instead of most frequent values.

The third law of thermodynamics. The problem of determining whether a chemical reaction or physical process will take place can be reduced to that of deciding whether or not the free energy of the whole system will be increased or decreased by the occurrence of the reaction. Consequently, the determination of these free energies (or rather of free energy differences) is of great importance. From the equation of definition $F = U - TS$, it follows at once that $\Delta F = \Delta U - T\Delta S$ for an isothermal change. Now ΔU is the heat of the reaction at constant volume, and ΔS is the change in entropy. Thus the sign of ΔF and therefore the direction in which the process will occur, could be predicted if the heat of reaction and the entropy change were both known. We may eliminate ΔS by the relation

$$\partial \Delta F/\partial T = -\Delta S,$$

so obtaining the differential equation

$$\Delta F = \Delta U + T\,\partial \Delta F/\partial T.$$

This can be integrated by dividing by T^2, which gives it the form

$$\frac{\Delta F}{T^2} - \frac{1}{T}\frac{\partial \Delta F}{\partial T} = \frac{\Delta U}{T^2},$$

i.e.
$$\frac{\partial}{\partial T}\left[-\frac{\Delta F}{T}\right] = \frac{\Delta U}{T^2},$$

so that
$$-\frac{\Delta F}{T} = A + \int \frac{\Delta U.dT}{T^2},$$

or finally
$$\Delta F = -AT - T\int \frac{\Delta U.dT}{T^2}$$

$$= -AT - T\int \frac{QdT}{T^2},$$

if we write Q for the heat of reaction. Now, regarded as a function of T, the heat of reaction, from the first law of thermodynamics, obeys the equation

$$\frac{dQ}{dT} = C_v' - C_v'',$$

where C_v is the specific heat at constant volume and the primes refer to the reactants and products respectively. We thus obtain ΔF in the form

$$-AT - T\int\left[Q_0 + \int_0^T (C_v' - C_v'')\,dT\right] dT/T^2 \qquad \text{.....(1)},$$

where now Q_0 is the heat of reaction at the absolute zero of temperature. Since

$$\Delta F = \Delta U + T\frac{\partial \Delta F}{\partial T},$$

this yields the known equation

$$\Delta U = \left[Q_0 + \int_0^T (C_v' - C_v'')\,dT\right].$$

Equation (1) is an expression for ΔF which can be evaluated if we know Q at any one temperature, the specific heats as functions of temperature, and the constant A. This last remark embodies a difficulty which chemists have felt for over a century. The constant A can only be determined by measuring the work obtained from the reaction (most easily by e.m.f. measurements) at some temperature; in other words, one point on the free-energy curve must be determined experimentally. Now from the time of Le Chatelier and Kelvin there has been an instinctive belief that it should be possible from purely thermal measurements to determine ΔF, i.e. in which direction a reaction will proceed. The analysis just given shows that this problem is equivalent to that of determining A, and to Nernst[15] falls the honour of having first seen clearly how this could be done.

From the fundamental equation

$$\Delta F = \Delta U - T\Delta S$$

it follows that at absolute zero (since ΔS is not infinite),

$$\Delta F = \Delta U.$$

(Parenthetically, we may remark that it is because we live relatively so near to absolute zero that this is often approximately true, and so the sign of the heat of reaction often agrees with that of ΔF.) Nernst's hypothesis, which dates back to about 1906, was that for condensed systems (liquids or solids) these quantities ΔF and ΔU approach their common value asymptotically, so that $\partial\Delta F/\partial T$ and $\partial\Delta U/\partial T$ both tend to zero as T tends to zero. Using the above expressions, this double assumption shows that

$$-A-\int\frac{\left[Q_0+\int_0^T(C_v'-C_v'')\,dT\right]dT}{T^2}-\left[\frac{Q_0+\int_0^T(C_v'-C_v'')\,dT}{T}\right]=0$$

and $\int_0^T(C_v'-C_v'')\,dT=0$, as T tends to zero.

On substitution from the second equation in the first, the latter becomes

$$-A-\int\frac{Q_0\,dT}{T^2}-\frac{Q_0}{T}=0, \quad \text{i.e. } A=0.$$

Thus Nernst's hypothesis solves the problem of determining the constant A in the simplest way possible, viz. by making it vanish.

We find therefore that

$$\Delta F=-T\int\frac{\Delta U.dT}{T^2},$$

and since $\Delta F=\Delta U-T\Delta S$, we have also

$$\Delta U-T\Delta S=-T\int\frac{\Delta U.dT}{T^2}=-T\left[-\frac{\Delta U}{T}+\int\frac{d\Delta U}{dT}\frac{dT}{T}\right].$$

Hence it also follows from Nernst's hypothesis that ΔS tends to zero with T, since the factor $d\Delta U/dT$ in the integrand does so. This conclusion that $\Delta S=0$ shows, moreover, not only that

$$\int(C_v'-C_v'')\,dT=0,$$

but that $(C_v' - C_v'')$ must tend to zero more rapidly than T itself, since

$$\Delta S = \int \frac{(C_v' - C_v'')}{T} \, dT.$$

It is thus in agreement with later experimental knowledge (much of it obtained at the instance of Nernst himself) and with the T^3 law (which follows from Debye's theory of specific heats).

Later, in 1914, Nernst considered the question whether the law applies also to gases[16], and concluded that it does. He did not, however, make any direct use of this last assumption, since, as he points out, it is difficult in the extreme to obtain a gas at very low temperatures without liquefaction supervening. When absolute values for entropy, internal energy or free energy are required, he prefers to start with the substances in the solid form at absolute zero, transforming to gas at a higher temperature. In this way, as is well known, the vapour-pressure curve is introduced into the theory, and the constant which occurs in this equation is the familiar Nernst's chemical constant. We do not propose to enter further into this question here.

The above exposition has been given at considerable length, although it hardly represents very recent work, partly because Nernst's own exposition (and most textbooks) make the supposition that the specific heats can be expanded in a simple power series, and partly because it is desirable that the actual consequences of Nernst's own assumption should be clearly traced out and collected together. It is important to do this, since recent thought has been questioning the truth of the theorem, and before giving a summary of these developments it is well to be certain what is involved in the theorem and what is due to later extensions of it.

Returning then to the fact that $\Delta S = 0$ at absolute zero, we note particularly that Nernst's theorem does not involve the assumption that S itself is zero either for the products or reactants. It is true that it is consistent with such an assumption, and indeed the postulate that $S = 0$ for all condensed substances at absolute zero was introduced by Planck[17], and adopted by Nernst in later expositions of his work. Planck reached the conclusion on statistical-mechanical grounds analogous to those which we have explained in the section on statistical mechanics. It involves the assumption that C_v itself must be an infinitesimal of higher order than T, and is not in contradiction to the assumption that $\partial F/\partial T$ and $\partial U/\partial T$ tend asymptotically to their common value zero at absolute zero.

The difficulty of the additional hypothesis can be seen when we remember that an irreversible process is accompanied by an increase of entropy, and since amorphous substances are believed to be unstable relative to the crystalline forms of the same substances, it would seem natural to expect that S would be lower for the crystal than for the vitreous solid, at all temperatures down to absolute zero. From the point of view of entropy as a measure of lack of order, the same conclusion is irresistible in the light of our knowledge of the constitution of the two states. Now it is impossible for both S and dS/dT to vanish for both forms under this restriction, and doubts were therefore felt as to the validity of this expression of Nernst's

theorem. A careful study of the specific heats of crystalline and amorphous forms of the same substances[18] was undertaken by Gibson and Giauque, and they found that the entropy difference did in fact exist. The "third law" of thermodynamics was therefore restricted to crystalline materials, and vitreous solids and liquids excluded from its immediate scope.

But after this stage had lasted for a very few years, new advances in theory again raised doubts as to the validity of the theorem even for all crystalline materials. Evidence was accumulating to show that in a crystal, heat motion does not cease even at absolute zero—that there is a zero-point energy could not, on the recognition of the truth of the Fermi-Dirac statistics, be denied; for these assert that no two atoms can have all their quantum numbers the same, and hence not more than two atoms can be in the lowest state of energy (allowing for two states of spin).

In recent years, the aspect of the third law which has received by far the greatest attention—and been the centre of most controversy[19]—is this one now under discussion, which states that entropy and not merely entropy-difference vanishes at absolute zero. The difficulty is to find out *a priori* the substances, or groups of substances, for which it is true. This question can now be satisfactorily answered, but unfortunately the answer is such as to show that pure thermodynamics could not have solved the question unaided. We have seen above how entropy in statistical thermodynamics is intimately related to the number of complexions, and hence it is easy to see qualitatively that the quantum-mechanical concept of degenerate states is bound up with the question since it affects the number of complexions. To remind the reader of this concept, we may refer to the fact that an energy level in an atom is said to be degenerate, when the level, although having only one energy, is yet really formed by the fusion of several levels of equal energy, which can be separated experimentally by a magnetic field or other means; this has a different effect on them, so that the levels now differ slightly in energy, and the spectral lines involving this level become multiple in the Zeeman effect. Such multiple levels, which of course are still present at the lowest temperatures, add to the complexions, thus increasing their number and with it the entropy. Similarly, a solid solution, say NaCl and NaBr, has more complexions than either NaCl or NaBr, since two atoms in the same states may exchange roles and so increase the number of complexions. The details of an extremely interesting experimental test of this point in the theory were given in volume 1 of these reports[20]; it confirms the theory as accurately as can be expected. We may note that to calculate the theoretical zero-point entropy, it is not necessary to classify, but only to enumerate the levels. The utilization of group theory for this is considered by Wilson[21].

Thus the answer to the question involves of necessity quantities outside the realm of pure thermodynamics, which can only be determined either by quantum mechanics or by spectroscopic or allied investigations. In fact, not only must supercooled liquids be excluded from the theorem, but all condensed phases containing degenerate states; the systems which are left, to which the theorem is directly applicable, have been well described by Eastman and Milner[20] as "phases of sharply specifiable energy".

Curiously enough, indeed, the science of spectroscopy, after annexing the entire domain of atomic theory, and proceeding to the peaceful occupation of the realm of chemical statics, is now actively engaged in the subjugation of our own province. No longer are the thermodynamic functions determined by thermal experiments, measurements of equilibrium constants and e.m.fs. of cells. A more accurate method is to deduce them from the energy levels revealed with precision by spectroscopy, and this is certainly a case where we should welcome the conqueror. Not only is the accuracy of these new methods greater than that of the old, but they are a step in the direction which is the avowed aim of science—to correlate different phenomena, and to enable the behaviour of systems to be predicted (numerically) in conditions to which they have not yet been experimentally subjected.

Calculations of thermodynamic functions. We remarked earlier that it is now possible to calculate thermodynamic functions from spectroscopic data, and, at the risk of making this article unduly long, a short account of the underlying ideas in this field seems desirable.

The analysis of band spectra shows that the energy of a molecule varies with the energy levels in which the electrons are to be found, and depends also on the nuclear vibrations and on the rotation; all three types of energy are quantized, and to a first approximation, the total energy may be expressed as the sum of three terms, one due to each effect. Thus the rotational energy, for a symmetrical top[22], is of the form

$$h^2j(j+1)/8\pi^2A+(1/C-1/A)\,h^2k^2/8\pi^2,$$

where j and k are the two relevant quantum numbers (subject of course to appropriate selection rules), A and C are the corresponding moments of inertia, and h is Planck's constant. The vibrational levels for a simple harmonic oscillator[23] would be

$$(n+\tfrac{1}{2})\,h\sqrt{k}/2\pi\sqrt{M},$$

where k is the restoring force for unit displacement and M the reduced mass. Thus for a molecule containing n atoms and therefore $3n$ degrees of freedom, of which 6 are translational and rotational, leaving $(3n-6)$ of vibrational freedom, there would be $(3n-6)$ terms each of the above form. Actually, however, the fact that the potential energy is not a linear function of distance causes the motion to deviate from the simple harmonic type, so that the energy is empirically expressed[23] (by analysis of the band spectra) in a series involving products and powers of the quantum numbers n instead of the above linear form. In the same way, the expression for rotational energy has to be modified to include squares of j and k, the current explanation of the reason being the stretching of the molecule by centrifugal force when it is rotating.

To apply the information obtained from spectroscopy to solving the problem of calculating thermodynamic functions requires an application of statistical mechanics, simple and direct in its principles, but complex in practice. We have seen that the absolute entropy of a system is given by $k \log W$, where W is the number of complexions; in the case where we are interested in the temperature-dependence of the entropy, we must of course treat molecules in different energy-

levels as being different in kind. The distribution, as always, is governed by the Boltzmann's factors $e^{-E/kT}$, where E is the energy of any one kind of molecule, k is Boltzmann's constant R/N and T is absolute temperature, R being the ordinary gas constant and N the number of molecules in one gram-molecule.

The outcome of a detailed but not difficult calculation is to show that

$$U = -Nk\,Q'/Q,$$
$$F = T\,[k \log N! - Nk \log Q],$$
$$S = Nk\,[\log Q - Q'/QT] - k \log N!.$$

Here U, F, S are the internal energy, Helmholtz free energy and entropy of one mol of the gas, whilst $Q = \Sigma e^{-E_r/kT}$, the summation is taken over all the different energy-levels E_r, multiple levels being counted the necessary number of times, and Q' is dQ/dT. Q is known as the partition function, or in Germany as the state-sum (*Zustandsumme*). The specific heat C_v may of course be calculated from the relation $C_v = dU/dT$, and the enthalpy I from $I = U + pv$, from which we have the further step $C_p = dI/dT$. The translational energies must naturally be included, although they have not been shown explicitly above on account of their relatively simple and familiar form, the contribution to Q from each degree of translational freedom being $\Sigma e^{-n^2h^2/8ml^2kT}$, where m is its mass and l^3 the volume.

From this relatively simple method, there has developed an immense literature[24] devoted to the mathematical problems encountered in the practical performance of the summations which the formulae direct us to perform. For this subject, which in any case is a matter for a few specialists, we have no space. A good summary is to be found in the 1936 paper by L. S. Kassel[24].

It is also to be noted that the degeneracy contribution to the entropy due to nuclear spin is of no direct interest in calculating the entropy differences associated with physical or chemical changes. In the same way, if more than one isotope is present, there is an entropy of mixing. In consequence, it is becoming common to omit this contribution from the calculated entropy, the resulting values being referred to as "practical" entropies, or "entropies on the practical scale".

§ 2. APPLICATIONS OF THERMODYNAMICS

It is doubtful if any one person could be familiar with all the fields of application of thermodynamics, and even if he were it would be ridiculous to attempt to survey them in a single article. All that can be attempted is, as in the case of § 1, to direct attention to some of the parts of the subject where most change has taken place in recent years (not necessarily only the last two or three), giving more or less detail according as other collected accounts are less or more easily accessible.

Probably the most striking advances, at any rate to the physicist, are in connexion with low-temperature physics and the demagnetization technique for the attainment of extremely low temperatures. This matter, has, however, received so much attention in the present series of *Reports* (volume **1**, McLennan, p. 198; volume **3**, Powell, p. 151; volume **3**, A. H. Wilson, p. 262) that nothing will be added here.

A rather more general discussion of the thermodynamics of magnetism has been given by Stoner[25], and the peculiar relations between the two forms of liquid helium[26] have been the inspiration of a good deal of writing[27], not always based on very good thermodynamics. They will be mentioned again below. The theory of solutions is somewhat older than these matters, but there seems to be no summary which is at the same time comprehensive and comprehensible, and a short account is therefore appended here.

A variety of more chemical applications has of course been made, but the more important of these have been discussed by different authors in *Chemical Reviews* and in the *Annual Reports of the Chemical Society*, so that they also may be ignored here.

The thermodynamics of magnetism. The principles of thermodynamics apply equally whether there is magnetization or not, but in considering magnetic problems it is naturally convenient to modify the equations so as to show explicitly the magnetic terms. Thus when a portion of material undergoes any changes, the increase in its internal energy U is, as always, the difference between the heat dQ added to it and the work done by it, but we now write the latter term as $(PdV - HdM)$ where P, V are pressure and volume, whilst H is the field-strength (assumed constant during the change) and M the magnetic moment of the material. Thus HdM is the work done by the field, or $-HdM$ is that done magnetically by the material. In the common case where dQ is expressible as TdS, the equation then becomes

$$dU = TdS - PdV + HdM,$$

and by the first law of thermodynamics U is a complete differential.

To take account of cases where H varies, it is also convenient to define a second quantity

$$U_2 = U - HM,$$

which also is a function of state only, and for symmetry, to denote U by the symbol U_1. In the same way, each of the quantities enthalpy, free energy and Gibbs's function has a pair of analogues $U - pv$, $U - TS$ and $U - TS + pv$, in one of which $U = U_1$ and in the other $U = U_2$.

Each of the eight functions is a function of state and they give rise to relations of the type $(\partial T/\partial M)_S = (\partial H/\partial S)_M$ which are true either at constant volume or at constant pressure. That is, $(\partial T/\partial M)_{S,P} = (\partial H/\partial S)_{M,P}$ and $(\partial T/\partial M)_{S,V} = (\partial H/\partial S)_{M,V}$, though (as Stoner emphasizes) it does not follow that $(\partial T/\partial M)_{S,P} = (\partial T/\partial M)_{S,V}$. The complete set is

$$(\partial T/\partial M)_S = (\partial H/\partial S)_M,$$
$$(\partial T/\partial H)_S = -(\partial M/\partial S)_H,$$
$$(\partial S/\partial M)_T = -(\partial H/\partial T)_M,$$
$$(\partial S/\partial H)_T = (\partial M/\partial T)_H,$$

where T, S, M and H have their usual meanings.

Stoner then points out that if we work either at constant pressure or at constant volume, the terms PdV or VdP may be omitted from the definitions of U_1 and U_2,

which consequently become formally identical with the equations of non-magnetic thermodynamics, with H in place of P and $-M$ in place of V. Under the conditions stated this is exact and not merely an approximation.

If the equations are applied to unit mass of material, the moment M becomes σ, the specific intensity of magnetization, and the equation of state is $\sigma = f(H, T)$. The detailed treatment of the results which follow when different forms are assumed for f must be sought in the original paper, since they fall within the subject of magnetism rather than that of thermodynamics.

Helium at low temperatures. The early attempts to solidify helium, by further decrease of the temperature below that which it liquefies, were unsuccessful, and it was only by increase of pressure that the result was attained. It has since been found that there is a second form of liquid helium, and that the original failure was due to the peculiar shape of the phase diagram. The familiar diagram for most substances, with pressure plotted vertically and temperature horizontally, has three lines radiating from the triple point, one (the hoar-frost line) extending downwards to the left, one upwards to the right to separate liquid from vapour, and a nearly vertical one between liquid and solid. With the intervention of a second liquid, the solid-vapour line opens out into two, with a channel between, occupied by liquid II. Liquid I lies to the right of this, so that the decrease of temperature from liquid I leads, not to solid, but to liquid II. Such a phase diagram is not uncommon in other systems, but thermodynamically, a point of interest lies in the fact that for helium there appears to be no latent heat of transformation for the change from liquid I to liquid II. The question has thus arisen whether they should be classified as different phases or as a single phase. Of course, to agree with the phase rule, they must be considered separate, if there is a single line on the $\{p, T\}$ diagram separating the regions occupied by them, but in certain other respects it is preferable to regard them as a single phase. The relationship has been named a phase change of the second order, and has led already to a considerable literature[27]. The property, which is at present of no direct thermodynamic interest, possessed by helium II, of having a thermal conductivity 200 times that of copper at room temperature, is probably well known to most readers, but may be new to a few and is on that account mentioned here.

Theories of solutions. Whereas the earlier theories of electrolytes assumed that the degree of dissociation into ions varied with concentration, and could be deduced by measurements of conductivity, vapour pressure or freezing-point, the more recent theories (for strong electrolytes) all assume, following Bjerrum's suggestion[28], that dissociation is complete, or practically so, over the whole range with which the theory is concerned. The effect of variations in concentration are then to be explained as due to the atmosphere of solvent which clusters around each ion, and to the effect of electrical interionic forces.

Qualitatively, it is easy to see that if the number of ions does not change with concentration, then in a dilute solution the ions are farther apart than in a concentrated one, and consequently the effect of electrical attractions is less marked in the former case. The fact that these interactions must be of effect had of course

been realized for a long time, but it was only about 1923 that Debye and Hückel succeeded in finding a mathematical approximation which is sufficiently representative of physical reality to give useful results[29]. Even this theory, it must be remembered, applies only to quite dilute solutions, and it is perhaps worth emphasizing that no attempt has ever yet been made to account for the properties of concentrated (electrolytic or non-electrolytic) solutions. The sort of thing we are apt to forget is that if two liquids A and B are mutually soluble, but not miscible in all proportions, then a mixture of the two separates into two phases S_1 and S_2 which must of necessity have the same vapour pressure for each component, since they are in equilibrium. If S_1 is a dilute solution of A in B, then the vapour pressure of B over it is lowered slightly from that of pure B in the usual way. It follows that the vapour pressure of B over S_2, the dilute solution of B in A, i.e. the very concentrated solution of A in B, is still just below that of pure B, although there is so little of it present.

But to return to the Debye-Hückel theory, the first assumption made is that the partial free energy (Gibbs's function) is the sum of a part due to thermal motions, and a part due to electrical actions, though in reading their paper it must be remembered that they use a function ϕ which is really Planck's potential and is equal to Gibbs's function divided by T, the absolute temperature. In calculating the electrical energy, the thermal motion intervenes, in that it affects the number of positive and negative ions simultaneously present in unit volume. If there are n_i ions per unit volume of the ith kind, with valency z_i, and if ν_i is the number of these ions produced by the dissociation of one molecule, then the density of the ions of the ith kind, k being Boltzmann's constant, is

$$\rho_i = n_i e^{z_i(\epsilon\psi/kT)},$$

where ψ is the potential and ϵ the electronic charge. This application of Boltzmann's factors is open to a certain degree of criticism in that it neglects the effect of fluctuation, but accepting it, it is to be combined with Poisson's equation

$$\nabla^2\psi = -\frac{\psi\pi\rho_i}{D},$$

where D is the dielectric constant of the solution. The elimination of ρ_i, and neglect of certain small terms, leads to the equation

$$\nabla^2\psi = \psi/b^2,$$

where b has the dimensions of a length and is defined by

$$\frac{1}{b^2} = \frac{4\pi\epsilon^2}{kTD}\,\Sigma n_i z_i^2.$$

We may note further that $\Sigma n_i z_i^2$ is usually known as the ionic strength I and was proposed as a better basis of measurement than concentrations themselves, by Lewis and Randall, who originally introduced the quantity on empirical grounds[30]. A second quantity with the dimensions of length is required in the theory, viz. the mean distance a to within which an ion can approach another; this is larger than the inter-atomic distances in polar crystals, because of the atmosphere of solvent

molecules, but functions in the theory as a sort of ionic radius. The potential energy is thus

$$-\frac{\Sigma n_i z_i^2 \epsilon^2}{2D\,(b+a_i)},$$

whence the thermodynamic potential is obtained and hence (by pure thermodynamics) such quantities as activity, depression of freezing point and osmotic coefficient.

Guggenheim[31] and others make a distinction between the rational osmotic coefficient

$$g=\frac{\mu-\mu^0}{RT\log(1-\Sigma x_i)}$$

and the practical coefficient $\dfrac{-A\,(\mu-\mu^0)}{RT\Sigma m_i}$,

where μ is the chemical potential (μ^0 under standard conditions) of the solvent, and m_i is the molality (i.e. number of mols per kg. of solvent), A being a definite constant dependent on the actual solvent, with the value 55·5 for water. x_i is the molar fraction.

The (rational) osmotic coefficient g, for water as solvent, is found to be given by

$$1+\frac{0{\cdot}3741I^{\frac{1}{2}}\,(x_w+\Sigma x_i)\,\sigma\,(aI^{\frac{1}{2}}/3{\cdot}08)\Sigma x_i z_i^2}{(x_w v_w+\Sigma x_i v_i)\log(1-\Sigma x_i)},$$

which reduces for dilute solutions to

$$g=1-\frac{0{\cdot}3741I^{\frac{1}{2}}\Sigma n_i z_i^2\sigma\,(aI^{\frac{1}{2}}/3{\cdot}08)}{\Sigma n_i},$$

where I is the ionic strength, x_i is the mol fraction of the ith ion, x_w that of the solvent, v_i its partial molecular volume and $\sigma(\lambda)$ is the function

$$\frac{3}{\lambda^3}\left[(1+\lambda)-\frac{1}{(1+\lambda)}-2\log(1+\lambda)\right].$$

In very dilute solutions, this reduces further to

$$g=1-0{\cdot}270w\sqrt{(\nu c)},$$

where c is the molal concentration and $\nu=\Sigma\nu_i$. For other solvents, the approximate equation is

$$g=1-\frac{we^2}{6DkT}\sqrt{\left(\frac{4\pi e^2 n\Sigma\nu_i}{DkT}\right)}.$$

This equation contains a valency factor

$$w=(\Sigma\nu_i z_i^2/\Sigma\nu_i)^{\frac{3}{2}},$$

which takes in common cases the values shown in table 2.

Table 2. Valency factors

Ions of opposite sign	Ions of first sign (per mol.)		
	1	2	3
1	1	2·83	5·2
2	2·83	8	16·6
3	5·2	16·6	27

The outstanding result of the Debye-Hückel theory, which has received a great deal of experimental investigation[32], is the deduction that the osmotic coefficient is (at sufficiently great dilution) a linear function of the square root of the concentration and of the cube of the valency z. This square-root law also applies to the viscosity, the partial molal volume and other derivatives of the free energy.

Actually the results for different solvents do not all agree with this equation, and Bronsted[33] has proposed the empirical substitute

$$g = 1 - \alpha' c^{\frac{1}{2}} - \beta c,$$

where β is a constant characteristic of the electrolyte and α' a constant determined by the valencies. This formula is empirically extremely satisfactory, and has the advantage that at low concentrations the α' term is the more important, so that the theoretical equation applies (since α' is actually proportional to Z^2), whereas at higher concentrations the β term becomes more prominent and accounts for the specific differences between electrolytes of the same valency.

Attempts to refine the theory so as to account for these specific differences have been made by different authors, notable contributions being due at different periods to Guggenheim[34], Onsager[35], Falkenhagen[36], Fuoss[37] and others[38]. Falkenhagen has considered in particular the viscosity[36], and it is found that a law

$$\eta = \eta_0 (1 + A\sqrt{c} + Bc)$$

is satisfactory, when A is given the theoretical value, so that effectively the only adjustable constant is B.

Turning now to the other property which differentiates electrolytes from non-electrolytes, we find that the equivalent conductance of an ion as calculated by Onsager[35] is again a linear function of the concentration. At a fixed concentration it varies in a complicated way with valency and with temperature. Physically, there are two independent effects taken into account: (1) when a field is applied, the ion is dragged ahead of its atmosphere which thus exerts an attractive force tending to retard the ion, and (2) there is a counter flow of solvent molecules, which again retards the ion. These effects vary at different rates with temperature and valency (and in fact with the viscosity of the solvent), so accounting for the relatively complex theoretical variation of conductivity with these variables.

Finally, we may add a warning that the very satisfactory agreement between this theory and the experimental results at high dilution cannot be taken as absolute proof of the fundamental truth of the theory. Gucker[39] reviewing the experimental evidence found that various properties of aqueous solutions of organic materials also varied linearly with the square root of the concentration. It is clear that in this case some utterly different mechanism must be postulated to account for the results, and it is therefore just remotely possible that it would apply also in the case of electrolytes, and that the current explanation is the wrong one. This hardly seems likely, however, since not only the form, but the actual values of the constants are correctly predicted, and since also the electrical conductivity follows the same law.

REFERENCES

(1) LABY, T. H. and HERCUS, E. O. *Philos. Trans.* A, **227**, 63 (1927).
(2) LABY, T. H. *Proc. phys. Soc.* **38**, 169 (1925–6).
(3) MAY, A. N. *Progr. Rep. phys. Soc.* **3**, 89 (1936).
(4) ROSSINI, F. D. *Chem. Rev.* **18**, 233 (1936).
ROTH, W. A. *Z. Elektrochem.* **41**, 112 (1935).
(5) International Institute of Refrigeration, Unpublished report.
(6) ROBERTS, J. K. *Heat and Thermodynamics*, 1st ed. p. 316.
PLANCK, M. *Thermodynamics*, Chapters 4 and 5.
NERNST, W. *Theoretical Chemistry* (1923 edition), p. 521.
(7) LEWIS, G. N. *Proc. Amer. Acad. Arts. Sci.* **43**, 259 (1907).
(8) RICE, J. *Statistical Mechanics*, p. 29.
(9) BRILLOUIN, L. *Ann. Phys.*, Paris, **7**, 315 (1927).
(10) JEANS, J. H. *Dynamical Theory of Gases.*
FOWLER, R. H. *Statistical Mechanics.*
TOLMAN, R. C. *Statistical Mechanics.*
GIBBS, J. W. *Statistical Mechanics.*
(11) This is Liouville's theorem. See ref. (10), and particularly Jeans, p. 73.
(12) BOLTZMANN, L. *Vorlesungen über Gastheorie* (1877 ed.), p. 43.
(13) VON NEUMANN, J. *Proc. nat. Acad. Sci., Wash.*, **18**, 70 (1932).
TOLMAN, R. C. *Statistical Mechanics*, p. 40.
SCHRÖDINGER, E. *S.B. preuss. Akad. Wiss.* p. 434 (1925).
PLANCK, M. *Z. Phys.* **35**, 155 (1925).
(14) DARWIN, C. G. and FOWLER, R. H. *Phil. Mag.* **44**, 450 and 823 (1922).
FOWLER, R. H. *Proc. roy. Soc.* A, **113**, 432 (1926).
(15) NERNST, W. *Nachr. Ges. Wiss. Göttingen*, **1** (1906).
NERNST, W. *S.B. preuss. Akad. Wiss.*, p. 933 (1906).
(16) NERNST, W. *Z. Elektrochem.* **20**, 357 (1914).
(17) PLANCK, M. *Phys. Z.* **13**, 165 (1912).
(18) GIBSON, G. E. and GIAUQUE, W. F. *J. Amer. chem. Soc.* **45**, 93 (1923).
(19) BRUNNER, R. *Z. Phys.* **100**, 584 (1936).
EASTMAN, E. D. *Chem. Rev.* **18**, 257 (1936).
(20) EASTMAN, E. D. and MILNER, R. T. *J. chem. Phys.* **1**, 444 (1933).
(21) WILSON, E. B. *J. chem. Phys.* **3**, 276 (1935).
(22) RUARK, A. E. and UREY, H. C. *Atoms, Molecules and Quanta*, p. 439.
(23) *Progr. Rep. phys. Soc.* **2**, 43 (1934) and **3**, 388 (1935).
(24) KASSEL, L. S. *J. chem. Phys.* **1**, 576 (1933).
KASSEL, L. S. *Chem. Rev.* **18**, 277 (1936).
ZEISE, H. *Z. Elektrochem.* **40**, 662 (1934).
KASSEL, L. S. *J. Amer. chem. Soc.* **55**, 1351 (1933); *Ann. Rep. chem. Soc.* **32**, 84 (1935).
(25) STONER, E. C. *Phil. Mag.* **19**, 565 (1935).
(26) SATTERLY, J. *Rev. mod. Phys.* **8**, 347 (1936).
VAN LAAR, J. J. *Proc. Acad. Sci. Amst.* **39**, 612 (1936).
(27) EHRENFEST, P. *Proc. Acad. Sci. Amst.* **36**, 153 (1933).
JUSTI, E. and VON LAUE, M. *Phys. Z.* **35**, 945 (1934).
DEHLINGER, U. *Z. phys. Chem.* **28**, B. 112 (1935).
(28) BJERRUM, N. *7th Int. Congr. Chem.* (1909).
(29) DEBYE, P. and HÜCKEL, E. *Phys. Z.* **24**, 185 (1923) and **24**, 305 (1923).
FOWLER, R. H. *Statistical Mechanics*, 1st ed., p. 314.
(30) LEWIS, G. N. and RANDALL, M. *J. Amer. chem. Soc.* **43**, 1112 (1921).
(31) GUGGENHEIM, E. A. *Phil. Mag.* **19**, 588 (1935).
(32) FALKENHAGEN, H. *Chem. Rev.* **13**, 1–139 (1933).
(33) BRONSTED, J. N. *J. Amer. chem. Soc.* **44**, 938 (1922).
(34) GUGGENHEIM, E. A. *Proc. roy. Soc.* A, **135**, 181 (1932) and ref. (31).

(35) Onsager, L. *Phys. Z.* **27**, 388 (1926) (Conductivity).
Onsager, L. *J. chem. Phys.* **2**, 599 (1934) (Conductivity).
Onsager, L. and Fuoss, R. M. *J. phys. Chem.* **36**, 2689 (1932).
Onsager, L. *Chem. Rev.* **13**, 73 (1933).
(36) Falkenhagen, H. and Fischer, W. *Phys. Z.* **33**, 941 (1932) and **34**, 593 (1933); *Z. Elektrochem.* **39**, 517 (1933).
Falkenhagen, H. and Vernon, E. L. *Phil. Mag.* **14**, 537 (1932) (Viscosity).
Falkenhagen, H. *Rev. mod. Phys.* **3**, 412 (1931) and ref. (32).
(37) Fuoss, R. M. *Chem. Rev.* **17**, 27 (1935).
Onsager, L. and Fuoss, R. M. Ref. (35).
(38) MacInnes, D. A., Shedlovsky, T. and Longsworth, L. G. *Chem. Rev.* **13**, 29 (1933).
Shedlovsky, T. *J. Amer. chem. Soc.* **54**, 1405 (1932).
(39) Gucker, F. T. *Chem. Rev.* **13**, 111 (1933).

REFRIGERATION

By EZER GRIFFITHS, D.Sc., F.R.S., F.Inst.P.
Physics Department, National Physical Laboratory, Teddington, Middlesex.

§ 1. INTRODUCTION

The term refrigeration is commonly applied to methods for the production and the maintenance of low temperatures suitable for the preservation of perishable foodstuffs. Such methods also find applications in other fields of engineering such as, for example, air conditioning.

So far as the mechanical side is concerned, progress has been continuous for the past 100 years, but the pace of the development has been intensified in the last 20 years on account of experience gained in the course of the enormous increase in the trade in imported fruit[1,2,3,4] and also of the work on the purely biological side which has resulted in a better understanding of the chemical processes taking place in the foodstuffs during storage.

Mechanical refrigeration had its birth in the invention of the vapour-compression machine by Jacob Perkins[5] in 1834. Dr Linde[6,7] placed the subject on a scientific basis. At that time he was Professor of Thermodynamics and turned his attention to the various heat-engine cycles which could be used for the efficient production of cold. He was also a very capable practical engineer, and the horizontal double-acting ammonia compressor[8] which he designed held the premier place for many years.

In the theoretical field of refrigeration Linde was followed by Mollier[9], who in 1896 introduced the first of the series of charts to bear his name. This chart was the {entropy, temperature} diagram. In 1904[10] he devised a chart which is known as the $\{\phi, i\}$ diagram, ϕ being the entropy and i the total heat or enthalpy. Nowadays this is the chart generally referred to as the Mollier diagram. It is noteworthy insomuch that Mollier had then realized the practical advantages of oblique co-ordinates. Another diagram described in the same paper is known as the $\{p, i\}$—the {pressure, total-heat} chart. Designers make considerable use of such charts, particularly when considering new refrigerants for use with the vapour-compression type of machine.

References

(1) Griffiths, Ezer and Griffiths, E. A. "The refrigerated transport of apples overseas." *5th Int. Congr. Refrig.* (1928).

(2) Pole-Evans and Griffiths, E. A. "The influence of refrigeration in the preservation of fruit; with special reference to South African export varieties." *4th Int. Congr. Refrig.* **1**, 945 (1924).

(3) Griffiths, E. A. "Precooling and transport of fruit in the Union of South Africa." *Sci. Bull. Dep. Agric. S. Afr.* No. 89.

(4) Griffiths, Ezer. "A survey of the developments in the refrigeration of foodstuffs in the British Empire since the Rome Congress." *6th Int. Congr. Refrig.* **4**, B, 391 (1932).

(5) EWING. "Perkins's machine of 1834." *The Mechanical Production of Cold*, p. 82 (1908). Camb. Univ. Press.
(6) LINDE. *Z. ges. Kälteindustr.* (28 January 1895).
(7) LINDE. *Z. Ver. dtsch. Ing.* (2 February 1895).
(8) LINDE. English Patent No. 1458 (1876).
(9) MOLLIER. *Z. ges. Kältenindustr.* **3** (1896).
(10) MOLLIER. *Z. Ver. dtsch. Ing.* **48**, 271 (1904).

§ 2. NEW REFRIGERANTS

For a long period ammonia, carbon dioxide and sulphur dioxide were the refrigerants[11,12] in general use. In the last decade a number of new refrigerants have been introduced. These have characteristics which render them particularly suitable for specific purposes: for example, in air conditioning it is desirable to have a refrigerant which is non-toxic in the event of leakage. Many of these new refrigerants are of the chlorofluoride type.

Thomas Midgley, the discoverer of these organic chlorofluoride refrigerants, has given a graphic account of his search on the occasion of the award to him of the Perkin Medal[13]. Incidentally he describes how he arrived at the conclusion that the boiling-point of carbon tetrafluoride should be about $-136°$ C., although the value quoted in the *International Critical Tables* was $-15°$ C. Subsequently it was found by experiment that the boiling-point was $-128°$ C.

The most commonly used of the new refrigerants is dichlorodifluoro-methane, CCl_2F_2, known in the trade as "Freon"[14–20], which is said to be non-toxic in concentrations up to at least 20 per cent by volume for a period of 2 hr. and consequently is favoured for air-conditioning plants. It is a colourless liquid boiling at $-30°$ C. under standard atmospheric pressure. It is sufficiently akin in thermodynamic properties to ammonia that it might be substituted for this gas in the standard form of compressor without radical change of design. In contact with flame and very hot surfaces the gas is decomposed with the formation of toxic products which are exceedingly irritating. The liquid and the vapour are non-inflammable and non-combustible. Whilst it is non-corrosive to all metals commonly used in refrigeration work it is not recommended for magnesium alloys especially in systems where water may be present.

Data as to the physical constants of this refrigerant, culminating in the formulation of the necessary tables of data giving its thermodynamic properties, will be found in the series of papers published by Buffington, Gilkey and others and also by the Kinetic Chemical, Inc., Wilmington, Del.

Another refrigerant introduced for use with air-conditioning plants is CH_2Cl_2 known as "Carrene". The refrigerant is well adapted for use in centrifugal machines of large capacities and where the operating pressures are relatively low. Another of the fluorine group—trichloro-monofluoro-methane, $CFCl_3$—is said to be suitable for use in turbo-compressors. Dichloro-tetra-fluoro-ethane, $C_2Cl_2F_4$, is used in one make of domestic refrigerators. Yet another of the fluorine derivatives, CF_3Cl, has a temperature-pressure relationship comparable to that of carbon dioxide.

Table 1 has been compiled by Carrier and Waterfill[21] to show the characteristics of some of the commonly used refrigerants with special reference to air conditioning.

Table 1 (Carrier and Waterfill)

Refrigerant	Chemical formula	Pressure (lb./in², gauge) P_1 35° F.	Pressure (lb./in², gauge) P_2 95° F.	Compression ratio $P_2 : P_1$	Lb. min. circulated per ton	Suction volume (cfm ton)	Hp ton (adiabatic compression)	Type of compressor	Practical hazard	Mechanical construction
Ammonia	NH_3	+ 51·5	+ 181·1	3·02	0·423	1·85	0·655	Positive	Toxic	Heavy; steel
Carbon dioxide	CO_2	+513·3	+1265*	2·46	3·73	0·588	1·34	Positive	High pressure	Extra heavy; steel
Sulphur dioxide	SO_2	+ 9·6	+ 62·9	3·2	1·43	4·59	0·64	Positive	Very toxic, corrosive	Copper, steel
Methyl chloride	CH_3Cl	+ 24·0	+ 95·4	2·85	1·33	3·38	0·66	Positive	Inflammable, toxic	Medium pressure; copper, steel
Carrene no. 1	CH_2Cl_2	− 11·77	− 2·71	4·06	1·48	31·2	0·630	Centrifugal	None	Vacuum; iron, brass
Carrene no. 2	$CFCl_3$	− 8·46	+ 6·9	3·44	2·93	17·85	0·650	Centrifugal	None	Vacuum; iron, brass
Freon	CCl_2F_2	+ 32·6	+ 107·6	2·59	3·83	3·30	0·670	Positive	None	Medium pressure; iron, brass
Water	H_2O	− 14·63	− 13·92	8·15	0·198	583·0	0·68	Centrifugal and jet	None	Vacuum; iron, brass, aluminium

* Head pressure for maximum cycle efficiency (dry compression).

Table 2, by Morse[22], is a comparison of a few well-known refrigerants.

Table 2 (Morse). Characteristics of refrigerants (standard ton conditions). 5° F. evaporator. 86° F. condenser

	Freon-12 CCl_2F_2	Carbon dioxide CO_2	Ammonia NH_3	Methyl chloride CH_3Cl	Sulphur dioxide SO_2
Boiling point (° F. at atmospheric pressure)	−21·7	−108·4	−28·0	−10·6	14·0
Absolute pressure (lb./in² at 86° F.)	107·9	1039·0	169·2	95·53	66·45
Absolute pressure (lb./in² at 5° F.)	26·51	334·4	34·27	20·89	11·81
Heat-content of saturated vapour in 5° F. evaporator (B.t.u./lb.)	78·79	102·14	613·35	195·3	183·49
Heat-content of liquid leaving 86° F. condenser (B.t.u./lb.)	27·72	45·45	138·9	46·6	42·12
Refrigerating effect (B.t.u./lb.)	51·07	56·69	474·45	148·7	141·37
Pounds of refrigerant per minute per ton of refrigeration	3·916	3·528	0·422	1·345	1·414
Ft³ of liquid per lb. at 86° F.	0·012	0·027	0·027	0·018	0·012
In³ of liquid per min. per ton of refrigeration	83·9	162·8	19·6	41·7	28·9
Ft³ of vapour per lb. at 5° F.	1·485	0·267	8·150	4·529	6·421
Ft³ of net piston displacement per min. per ton of refrigeration	5·815	0·943	3·436	6·091	9·084
B.t.u. refrigeration per ft³ piston displacement	34·4	212·0	58·2	32·8	22·0

One characteristic of the new refrigerants might be noted. They are miscible with lubricating oils, and advantage is taken of this to keep all heat-transfer surfaces free from excess oil by automatically flashing with refrigerant.

References

(11) Griffiths, Ezer and Awbery, J. H. "Small refrigerating plants and the thermodynamic properties of refrigerating liquids." *Proc. Brit. Ass. Refrig.* (March 1925). [References to literature.]

(12) McGovern. "Methyl chloride—A practical refrigerant measured by the theoretical ideal." *J. Refrig. Engng*, **34**, 29 (1937).

(13) Midgley. Award of Perkin Medal. *J. industr. Engng Chem.* (February 1937).

(14) Buffington and others. "The thermodynamic properties of 'Freon'":

(1) Buffington and Gilkey. "The equation of state of the superheated vapour." *J. industr. Engng Chem.* **23**, 254 (1931).

(2) Gilkey, Gerard and Bixler. "Vapour pressure." *J. industr. Engng Chem.* **23**, 364 (1931).

(3) Bichowsky and Gilkey. "Critical constants and orthobaric densities." *J. industr. Engng Chem.* **23**, 366 (1931).

(4) Buffington and Fleischer. "Specific heat of liquid and vapour and latent heat of vaporisation." *J. industr. Engng Chem.* **23**, 1290 (1931).

(5) Buffington and Gilkey. "Correlation checks and derived quantities." *J. industr. Engng Chem.* **23**, 1292 (1931).

(15) Thompson. "The technical aspects of Freon refrigerants." *Refrig. Engng*, p. 225 (April 1937).

(16) Hodsdon. "Freon chart." *Proc. Brit. Ass. Refrig.* **32** (1935–6).

(17) Brizzolara. "Ammonia and 'Freon'." *Refrig. Engng*, p. 330 (May 1937).

(18) Griffiths, Ezer. "The properties of refrigerants recently introduced." *Proc. Brit. Ass. Refrig.* **32** (1935–6).

(19) *Rep. U.S. Bur. Mines*, R.I. 3013, 3185.
(20) *Rep. U.S. Underwriters Lab.* M.H. 2375.
(21) CARRIER and WATERFILL. *7th Int. Congr. Refrig.* **3**, 718 (1937).
(22) MORSE. *7th Int. Congr. Refrig.* **3**, 438 (1937).

§ 3. INSULATING MATERIALS[23]

For the insulation of cold stores a variety of materials have been tried but cork still holds the premier place. It was discovered by accident that cork granules heated under pressure to such a temperature as to turn the colour brown adhered together to form a compact block. Slabs of cork by this process are nowadays manufactured on a large scale. The density of the resulting product may vary from about 7 to 12 lb./ft.^{3} according to the mode of manufacture. It is found[27] that the lighter the slab the lower the thermal conductivity. For example, a 7 lb./ft.^{3} density cork had a thermal conductivity of 0·00007 g.-cal./cm.^{2}-sec. for 1 cm. thickness and 1° C. difference in temperature whereas a 12-lb./ft.^{3} material had a thermal conductivity coefficient of 0·00013.

Concrete with a cellular structure is also used in cold-storage construction. The cellular structure is produced by mixing with the cement a foam-producing liquid and blowing air into the mix. An alternate method of producing the cellular structure is to incorporate aluminium powder with the cement. This reacts with the alkali in the cement generating hydrogen and the bubbles give the cellular structure. Rubber and ebonite can be produced with a cellular structure by vulcanizing under pressure with a gas such as nitrogen and releasing the pressure whilst cooling.

Developments have also taken place in the air cavity form of insulation[28]. One form utilizes aluminium foil in the form of superimposed sheets separated by wooden distant pieces or kept apart by initially crumpling the aluminium. Another form uses very thin sheets of steel coated with a metal such as lead to reduce corrosion. The low thermal emissivity of the aluminium or lead is a factor which contributes to the insulating value of these partitions.

In the determination of the thermal conductivity of materials having a very low coefficient of thermal conductivity it is advisable to employ a guard ring method. In this an electrically heated plate with a coplanar guard ring is sandwiched between two slabs of the material under test, the whole being clamped between two cold plates. Many modifications of this form of apparatus have been described by the writer and others.

When the test specimen takes the form of a large slab measuring 3 by 3 ft. and of a thickness up to 12 in. another form of apparatus[29] is used. The hot plate is backed by a guard plate maintained at the same temperature and a system of side guard plates fitted in which the temperature gradient is the same as through the specimen under test.

The writer and Mr J. H. Awbery[30] have described tests in which the conductivity of a material is determined over the range $+10°$ to $-80°$ C. by utilizing solid carbon dioxide as the cooling medium and the calorimetric agent for the measurements of the heat leakage in through the walls of the container. A box is

made of the material under test. This is supported from the arm of a balance and the rate of loss of weight of the solid carbon dioxide measured. An essential feature is the provision made for keeping the carbon dioxide block in contact with the under surface of the lid.

References

(23) GRIFFITHS, EZER. *Thermal insulation.* Howard Lectures, Roy. Soc. Arts, 1933. [Additional references to literature will be found in this pamphlet. See also Report of the Insul. Comm. *Refrig. Engng*, **10**, 172 (1923/4).]
(24) GLAZEBROOK and GRIFFITHS, EZER. "The scientific study of heat insulators." *4th Int. Congr. Refrig.* (1924).
(25) GRIFFITHS, EZER. "Heat insulation in cold storage construction." *Struct. Engr*, p. 202 (May 1932).
(26) GRIFFITHS, E. A. "Cold storage construction." *Struct. Engr* (1932).
(27) GRIFFITHS, EZER. "Heat insulators." *Spec. Rep. Food Invest. Bd, Lond.*, No. 35.
(28) GRIFFITHS, EZER. "Apparatus for the study of heat insulating materials for use at low and high temperatures." *Chem. Engng Congr.* (1936).
(29) GRIFFITHS, EZER. "Thermal conductivity of insulating materials and methods of testing." *J. Instn Heat. Vent. Engrs* (March 1937).
(30) GRIFFITHS, EZER and AWBERY, J. H. "The thermal conductivity of materials used for solid CO_2 containers." *7th Int. Congr. Refrig.* **3**, 167 (1937).

§ 4. INSTRUMENTAL EQUIPMENT

Modern methods of cold storage require instruments for the measuring of the temperature, humidity and of the composition of the atmosphere.

Temperature measurement. Resistance thermometers are generally used for reading temperatures at a distance. Mercury-in-steel thermometers find a limited application.

The conditions prevailing in cold stores are particularly severe on resistance thermometers and trouble has frequently been encountered due to defective insulation caused by moisture condensation. It might be mentioned that the packing gland at the thermometer head is a source of potential trouble. In many quarters there are grave misgivings as to the reliability of the resistance method of measuring cold storage temperatures. Undoubtedly the selection of a temperature-measuring outfit for cold storage work calls for a careful scrutiny of the design to ensure freedom from the above sources of trouble. The usual form of resistance thermometer outfit is of the simple two lead form with an indicator of the direct deflexion type. To eliminate the necessity for adjustment of the battery current to a predetermined value cross coil instruments have been developed. The nul-deflexion type is also a solution of this problem.

A three-dial instrument has been designed by the writer[32] in which the temperature in the range 0 to 100° is read off from the numbers appearing beneath the dial windows. The switch gear is so arranged that a suitable resistance for each dial is interposed in the battery circuit to prevent the galvanometer receiving a violent deflexion.

Gas analysis. The carbon dioxide and the oxygen content of the atmosphere in a

stores has to be kept within prescribed limits. The composition does not vary rapidly as in flue gases, consequently periodic tests suffice. The gas burette form of apparatus can be used for carbon dioxide and oxygen and there are a variety of automatic instruments for carbon dioxide alone. The writer devised a pump type for CO_2 in which the absorption type of instrument is reduced to its simplest form. The hot-wire form of gas analysis instrument[34] has been installed on an extensive scale sometimes without a due appreciation of its limitations.

Sufficient consideration has not been given in cold storage work to other instruments for gas analysis based on the physical properties of the mixture. Instruments for the determination of CO_2 in flue gases have been designed which utilize the fact that CO_2 has a density 53 per cent greater than that of air. In one form[31] two centrifugal fans set up pressure heads that are directly proportional to the density of the media in which they rotate. The two fans are enclosed in chambers containing air and the gas mixture respectively. In each of the same chambers and facing the rotating fans are two fans capable of a small degree of rotation. The latter fans are coupled by a link such that the deflexion is the difference of the torques. The percentaged CO_2 is read directly from the deflexion of a pointer. Such instruments might well be tried for the determination of CO_2 in the atmosphere of gas stores. Another instrument[31] based on the physical properties of the gas mixture is the Rayleigh interferometer.

The writer has described a gas-analysis instrument based on the measurement of the velocity of sound in the mixture using a quartz crystal as source of sound and measuring the half wave-lengths of the standing wave system. For the determination of oxygen the writer has described a simple form of absorption instrument. In this the manipulation required is reduced to a minimum.

Hygrometry[35, 36]. The routine measurement of the relative humidity of the atmosphere in a cold stores can be effected by means of the wet and dry bulb hygrometer or the dew-point apparatus.

A convenient method of operating the wet and dry bulb hygrometer is to draw a stream of air from the stores past wet and dry bulb thermometers of the mercury in glass type inserted in the pipe line. Or the wet and dry bulb thermometer may take the form of a multiple element thermoelement carried at the end of a hose pipe through which a stream of air is drawn by a fan.

The writer and Mr J. H. Awbery[37] have studied the wet and dry bulb hygrometer at temperatures below 0° C. and published a set of tables for the reduction of the observations. In these experiments as standard of reference the gravimetric and dew-point methods were used. The apparatus for the gravimetric method took the form of a dry gas meter, a suction pump and silica gel absorption tubes. Relatively large volumes of the air from the stores had to be worked with to give a sufficient quantity of water for accurate weighing. The dew-point apparatus took the form of a flat silver disk which could be cooled by the circulation of paraffin oil through a spiral on the rear face. The oil was cooled by means of solid carbon dioxide. The temperature of the disk was obtained by means of a thermoelement pressed into contact with the surface.

References

(31) GRIFFITHS, E. A. and DAVIES, R. "Ocean transport of fruit under refrigeration from South Africa." *Sci. Bull. Dep. Agric. S. Afr.* No. 56 (1926). [Discussion of commercial forms of gas analysis apparatus.]
(32) GRIFFITHS, EZER. "Research in refrigeration." *Proc. Brit. Ass. Refrig.* (Oct. 1936).
(33) J. H. AWBERY and GRIFFITHS, EZER. "A study of a type of flowmeter depending on the pressure drop past a plug in the flow pipe." *Proc. Brit. Ass. Refrig.* **33**, 98 (1936–7).
(34) DAYNES. *Gas Analysis by Measurement of the Thermal Conductivity.* Camb. Univ. Press.
(35) GRIFFITHS, EZER and AWBERY, J. H. "The measurement of humidity in closed spaces." *Spec. Rep. Food Invest. Bd, Lond.*, No. 8, revised edition (1933).
(36) GRIFFITHS, EZER. "The measurement of humidity." *Sch. Sci. Rev.* (1932).
(37) AWBERY, J. H. and GRIFFITHS, EZER. "An investigation of the wet and dry bulb hygrometer at low temperatures." *Proc. phys. Soc.* **47**, 684 (1935).

§ 5. REFRIGERATION PROCESSES

Quick-freezing processes. One of the noteworthy advances in refrigerating engineering in recent years has been the development of machines for the rapid freezing[38,39] of meat, fish, vegetables and fruits. The meat or fish is prepared as fillets, packed into cartons and subjected to low temperatures. Fish when slowly frozen at -1 to $-5°$ C. and subsequently thawed has a spongy or woolly texture and the tissue fluids escape easily.

When the temperature of freezing is lowered, the amount of unfrozen tissue fluid of increasing salt concentration is diminished until the eutectic point of $-37°$ C. is reached. At these low temperatures, the tissues contain fine crystals, but after a period of storage in fluctuating temperatures the crystals grow. Hence the necessity for storage at constant low temperatures. Work has been carried out in Japan[40] on the size distribution of ice crystals and growth during cold storage.

In the quick freezing of soft fruits such as strawberries they are mixed with sugar and frozen. The effect of the sugar is to delay autolytic changes. Another method of preventing autolytic change is to destroy the enzymes by blanching or partially cooking before freezing; this method is applied to vegetables. It is found that when vegetables are to be preserved by quick freezing shorter blanching periods are permissible than would be the case were canning the process of preservation. This shortening of the period results in a frozen product which is of a good colour, flavour and texture.

The value of the freezing lies in its chemical and physical effects on the tissues[41]. It slows down the enzyme action in accordance with chemical kinetics. A lowering of 10 per cent in terms of absolute temperature will lessen the enzyme activity also by 10 per cent. The foodstuffs must not of course be kept for any length of time after thawing and in this respect resemble canned products. In one quick-freezing process the freezing of the packed product is effected by placing the cartons between two refrigerated flat metal surfaces such as belts or plates. The tendency nowadays is to use plates under controlled pressure. These refrigerated plates are controlled by a hydraulic piston and are made of an aluminium alloy having sinuous passages for the circulation of refrigerant which may be brine or ammonia. In operation the plates are cooled until they begin to frost up, and then the product to be frozen is

loaded between the plates. These are pressed on to the packages and the product left until it reaches a temperature of about −17·8° C. (0° F.). The relative freezing times for products of various thicknesses have been determined by tests using thermoelements. Rutov[42] has attempted a mathematical solution of the problem of freezing between flat surfaces. In another quick-freezing process the refrigerating effect is conveyed to the product through the medium of finely atomized salt brine at low temperatures. The transport and retail distribution of quick-frozen produce has necessitated the development of low-temperature rail transport. Heavily insulated refrigerator cars are employed. Some of the cars in use in the U.S.A. take 14,000 lb. of ice for cooling, and a mixture of ice and salt is used for temperatures below 0° C. It has been suggested that instead of ice and salt sodium chloride eutectic ice be used as this gives a temperature of −21° C. (−5·8° F.).

For road transport vans fitted with mechanically operated plants are used. These are operated during idle periods from the garage electric-power line. Dry ice is also used for road vans, but for rail transport it is too costly since long-distance hauls may occupy as long a period as 11 days. In the retail shops, the frozen product is kept in low-temperature refrigerators.

Refrigerated gas storage[43]. For a long period the usual method of dealing with fruit was to store at some predetermined temperature above freezing-point. In recent years, gas storage has come to the fore. In this method, the atmosphere in the stores has a moderately high carbon dioxide content and a low oxygen content. For example, in the storage of Bramley Seedlings a suitable atmosphere is obtained by allowing the carbon dioxide produced by the fruit to replace the oxygen in the air and then adjusting its concentration by controlled ventilation with fresh air. In this case the sum of the concentrations of oxygen and carbon dioxide always amounts to 21 per cent approximately. For some varieties of apples a low concentration of oxygen and of carbon dioxide is necessary. By thus controlling the composition of the atmosphere in the stores, the life of many varieties of apples can be greatly prolonged and the storage effected at temperatures sufficiently high to avoid the incidence of a functional disease known as low-temperature breakdown.

The design of gas stores involves difficult problems, as gas tightness must be obtained as well as ability in the structure to withstand temperature changes. In the gas stores designed by E. A. Griffiths[44] a welded steel structure is erected with girders capable of allowing for thermal expansion. The interior is lined with blocks of light concrete to give thermal insulation, and cooling is effected by grids of piping suspended from the ceiling.

Effect of volatile substances on fruit. One of the surprising discoveries of recent years has been that of the influence of minute traces of ethylene on the ripening of fruit. Apples, pears and certain other fruits produce a small amount of ethylene which may accumulate in parts per ten thousand of the atmosphere. The presence of this gas stimulates ripening, and when ripe fruit is kept in the same chamber as unripe fruit the rate of ripening of the unripe fruit is accelerated. The presence of minute traces of substances hitherto not recognized can cause heavy losses in the storage process. For example, it has been found that the injury known as scald can

be prevented by wrapping the apples in oiled paper. Presumably the oil absorbs some unknown substance or group of substances which accumulate on the surface of the fruit. The traces of volatile substances opens up a large field of research and calls for the development of a suitable technique[45] for producing controlled atmospheres.

Ice rinks. Artificial ice rinks have now been in existence for about 60 years. The Southport Glaciarium[46], opened in January 1879, had an ice floor produced by the circulation of glycerine and chloride of magnesium through flat pipes. The refrigerating plant was a 25 h.-p. sulphurous acid machine capable of manufacturing 24 tons of ice per day. The basic design of this form of ice rink construction was due to Prof. Gamgee. In various modified forms this system has been in use since that date. Recently Koopman[47] has put forward a design in which the ice layer is formed on the surface of a shallow bath of cold brine, the two liquids being separated by a thin film of a third substance such as paraffin.

Temperature distribution in refrigerated chambers. One of the problems encountered in the application of laboratory data to commercial work is to ascertain how nearly the conditions obtaining in practice approximate to those regarded as ideal. A ship with refrigerated holds is called upon to maintain specified temperatures in those holds and the extent of the variation in temperature can only be ascertained by laborious temperature surveys.

Much work on this subject has been carried out in recent years under the auspices of the Food Investigation Board, and a chamber has been installed at East Malling constructed as a section of the hold of a ship. The data obtained in the study of temperature distributions with various arrangements of the cooling system will be found in the *Annual Reports of the Food Investigation Board*. The main conclusion of all the work both on ships and in the experimental hold is that the system in which cold air is discharged into the chamber from ports in the ceiling is superior to all the others tested.

References

(38) Poole and Zarotschenzeff. "Advances in quick freezing during the past four years." *7th Int. Congr. Refrig.* **4**, 200 (1937).

(39) Barker and Morris. "The preservation of fruit and vegetables by freezing." *Leafl. Food Invest. Bd*, No. 2 (1932).

(40) Okuno. "Formation of ice crystals in the tissue of frozen fish." *7th Int. Congr. Refrig.* **4**, 290 (1937).

(41) Ezer Griffiths, Vickery and Holmes. "The freezing, storage and transport of New Zealand lamb." *Spec. Rep. Food Invest. Bd, Lond.*, No. 41.

(42) Rutov. "Thermal processes involved in the freezing of perishable foodstuffs." *7th Int. Congr. Refrig.* **4**, 211 (1936).

(43) Kidd and West. "Gas storage of fruit." *Spec. Rep. Food Invest. Bd, Lond.*, No. 30.

(44) Griffiths, E. A. "All steel gas store for fruit." *Proc. Brit. Ass. Refrig.* **32**, 103 (1935–6); *Industr. Gases* (Sept. 1936).

(45) Griffiths, Ezer and Awbery, J. H. "Production of air streams containing known small concentrations of essential oils." *Rep. Food Invest. Bd, Lond.*, p. 144 (1930).

(46) "Southport Glaciarium." *Liverpool and Southport Daily News* (11 January 1879).

(47) Koopman. "A new process for the realisation of artificial ice rinks." *7th Int. Congr. Refrig.* **3**, 562 (1937).

THE BEGINNINGS OF THE NEW QUANTUM THEORY

By H. T. FLINT, D.Sc., Ph.D.

§ 1. INTRODUCTION

Some twenty years ago the Physical Society published the *Report on Radiation and the Quantum Theory* written by Sir James Jeans (1914). The date of publication is a year later than that of Bohr's first application of the quantum theory to the physics of the atom. The great successes which were to develop from Bohr's revolutionary suggestions were to be reaped in the following decade, but it may be supposed that the success which attended even the work of one year was the stimulus to the writing and publication of the report.

It would appear that the success of Planck's theory in the study of the problem of radiation had not completely won over a large body of classical conservatives. The hope was still expressed that a return to classical methods might be possible, but the hope began to fade somewhat in the face of the triumphs of the application of quantum ideas in spectroscopy. The state of the quantum theory at the time of the publication of the report did, however, justify the placing of radiation first in the title.

If we were to attempt a report of a similar character to-day the order would have to be reversed, for modern theory has been more successfully developed in its application to the physics of the atom than to the theory of radiation.

The six years following the publication of the report were years of great activity and of triumphs for the quantum theory. In this period the Wilson-Sommerfeld conditions were discovered, and with their help a large field was successfully investigated in which old difficulties were satisfactorily settled and experiment stimulated.

Bohr's description of the relationship between the elements is one of the best known of the successes.

But at the end of this period, which lasted until about 1920, difficulties arose, and during the next five years serious doubts were entertained about the adequacy of the theory.

Bohr himself was one of the first to recognize the difficulties, and the leader of the revolution of a few years earlier began to stress its weak points. He often called attention to the defects revealed in the attempt to account for the anomalous Zeeman effect, and was of the opinion that an understanding of this phenomenon would point the way towards the next advance. It is now clear that he was guided by an unerring intuition in the right direction.

We will glance briefly at some of the difficulties which caused the decline of the quantum theory in the older form.

In this theory the energy of a rigid rotator of moment of inertia I may have values given by the equation

$$W_m = \frac{m^2 h^2}{8\pi^2 I},$$

where m is integral.

This formula was applied to account for rotation spectra as, for example, in the case of the halides of hydrogen. But in order to obtain agreement with experiment it was necessary to allow m to take the values $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, etc.

The integral numbers of the theory have to be replaced by the half-integer values.

This is a change which, from the point of view of the old quantum theory, is of the nature of an empirical correction. The energy formula corrected in this way gives the term of Deslandres and is of an importance in the application to systems like molecules equal to that of the Balmer term in its application to atomic spectra.

A difficulty arose in connexion with the anomalous Zeeman effect. When an atom is subject to a magnetic field of intensity H a change occurs in the energy values in the different levels of an amount given by the formula

$$\Delta W = mg \frac{he}{4\pi m_0 c} H.$$

m is a number known as the magnetic quantum number, while g is found from Landé's g-formula

$$g = 1 + \frac{j(j+1) + s(s+1) - l(l+1)}{2j(j+1)}.$$

The notation is that customarily adopted, l denoting the azimuthal quantum number reduced by unity, s the spin number and j, the inner quantum number, represents the total momentum.

The formula is verified by experiment to a surprising degree of accuracy, but the methods of the old quantum theory failed to establish it. According to these methods we expect to find the squares j^2, s^2 and l^2 instead of $j(j+1)$, etc.

This is the same substitution of squares of a quantum number λ by $\lambda(\lambda+1)$ as occurs in the case of a rotating molecule, for we can write the latter product in the form $\{(\lambda+\frac{1}{2})^2 - \frac{1}{4}\}$, and thus differences in energy values are obtained by replacing λ^2 by $(\lambda+\frac{1}{2})^2$.

A very remarkable difficulty arose in connexion with the explanation of the fine structure of spectral lines by means of Sommerfeld's theory. Early in the application of the old quantum methods Sommerfeld gave a satisfactory quantitative description of the fine structure of the hydrogen lines which remains accepted to-day. A general formula was developed for the calculation of energy levels on the assumption that the relativistic expression for the mass of the orbital electron must be applied. From this point of view it follows that the energy depends not merely upon the sum of the radial and azimuthal quantum numbers but also separately upon the latter. The characteristic of the problem is that it is concerned with the motion of a particle in a central field of force. It would be expected to

apply to the electron in the hydrogen atom, to the electron of the ionized helium atom and with a correction for the screening effect of other electrons to an electron concerned with the generation of X-ray spectra. The formula agrees well with experiment in these cases; it would therefore be concluded that fine structure can be accounted for by the relativity effect.

But Millikan and Bowen and simultaneously Landé called attention to the fact that the explanation offered in the cases in question could not apply to neutral alkali atoms. The same formula, however, gave satisfactory results. According to Millikan and Bowen the effect in this case is due to some effect other than that of the variation of mass with velocity. By a coincidence this effect led to the same energy formula. The difficulty was explained to some extent by the suggestion of Goudsmit and Uhlenbeck that the electron has a degree of freedom of motion about an axis.

The particular difficulty mentioned here cannot be regarded as indicating a complete breakdown of the old theory, but its great interest lies in the fact that in the new theory the problem is satisfactorily settled by Dirac's discovery of the need of a particular form of equation which forms the basis of the theory. His suggestion resembles that on which Sommerfeld based his theory of fine structure, for it rests on an appeal to the principle of relativity. But unlike that of Sommerfeld it solves the difficulty raised by Millikan, Bowen and Landé, and it does it in a way which throws light on the suggestion of Goudsmit and Uhlenbeck.

A more serious breakdown of the older quantum methods than that suggested by these examples is to be found in the study of the helium atom. If the principles are applied here we meet with failure. We learn from this that our results agree with experiment only in cases where the motion of a single electron is concerned. As soon as we pass to the atom next in order to the hydrogen atom, where the motion of two electrons interacting with one another has to be considered, failure draws attention to the inconsistency of the methods employed, an inconsistency which is masked by success in the simpler cases. The characteristic principle of the old theory is that the motion is described by classical mechanics and limitations foreign to this system of mechanics and, in some of their consequences, in contradiction to it are imposed.

It became clear in the declining years of the theory that a new idea was required before further advance was possible, and that a method was desirable which removed the inconsistencies as well as the difficulties from a system which, in spite of brilliant successes, could not be regarded as complete or even satisfactory.

§ 2. THE CORRESPONDENCE PRINCIPLE

Before leaving these references to the old quantum theory we shall turn with some detail to a principle which Bohr introduced in the early years of its development. It is a principle of a general character and was originally described by its author as a principle of correspondence between the wave theory and the quantum

theory. We shall see how it was applied and what service it rendered in the earlier form of the quantum theory, and in the development of the more recent theory we shall realize its importance on account of its influence upon the work of Heisenberg who widened and deepened the principle by expressing it in the language of the new quantum mechanics.

As a theory of radiation the old quantum theory had serious defects. It concerned itself with frequencies leaving aside questions concerning intensities and states of polarization. The classical theory of radiation had tried to give a deeper insight on all these points by relating the vibrations of a particle, which played the part of the exciter of radiation, with the frequencies of the radiation produced by it. If the vibratory motion of the particle was complicated in its character, it could be regarded as a summation of vibrations consisting of a fundamental frequency and of overtones. All this series of frequencies was expected to be present in the radiation produced, and the relative intensities of the particular frequencies were to be obtained from the amplitudes of the mechanical vibrations of the corresponding frequencies. The state of the mechanical vibration corresponded to the state of polarization of the radiation of the same frequency.

This classical theory is itself a principle of correspondence of mechanical motion with radiation, but what is meant by the principle of correspondence in the quantum theory is a correspondence between the classical and quantum theories.

The latter theory was soon developed to such a degree that in the point of accounting from the frequencies of spectral lines it reached a degree of success which made it independent, but it gave no information on the other two characteristics of radiation. This is where the principle of correspondence was of great and indispensable service. The results it gave with regard to intensities were unsatisfactory from a quantitative point of view, but with regard to polarization the information agrees with that provided by the recent work.

In order to see how Bohr applied the principle originally let us consider a charge e with an acceleration f. The electric and magnetic intensities of the field radiated by this accelerated charge are

$$E = H = ef/c^2R \qquad \text{.....(1)},$$

where e is measured in electrostatic units and R denotes the distance from the charge to the point where E and H are measured.

The factor controlling the calculation of the emission of radiation is thus the vector $e\mathbf{r}$, ef denoting $ed^2\mathbf{r}/dt^2$. This vector can be divided into its components ex, ey and ez, where x, y and z denote the coordinates of the charge from the position of rest as origin.

If we have to consider a collection of vibrating particles in an atom, the radiation field consists of the sum of a number of contributions and the important vector is

$$\mathbf{P} = \Sigma e\mathbf{r} \qquad \text{.....(2)},$$

which we can divide into the components

$$P_x = \Sigma ex, \quad P_y = \Sigma ey, \quad P_z = \Sigma ez \qquad \text{.....(3)}.$$

By a double differentiation of these we obtain quantities proportional to the total electric and magnetic components of intensity at a point distant from the atom, i.e. at a point such that R may be regarded as the same for all points of the atom.

Thus the problem is reduced to finding x, y and z by the application of mechanical principles to the complex structure of the atom.

We then express **P** in terms of its component vibrations by a Fourier expansion, and we assume that these describe the spectral lines.

In the case of complex systems the methods of general dynamics offer the most powerful method of proceeding with the determination of x, y and z as functions of the time. We shall, however, avoid a complicated representation of the principle and shall consider some simple examples to illustrate it.

The simplest of these is the oscillator considered by Planck in the early days of the quantum theory. Suppose that the displacement occurs in the x-direction, so that

$$x = a \cos 2\pi\omega t \qquad \ldots\ldots(4).$$

The electric moment

$$P = ex = b \cos 2\pi\omega t \qquad \ldots\ldots(5),$$

writing $b = ea$.

It follows from equation (1) that the electric and magnetic intensities are proportional to $4\pi^2\omega^2 b \cos 2\pi\omega t$, and according to the classical view we expect to find radiation of a single frequency ω as a result of the oscillations of the particle. The absence of terms in equation (5) containing other frequencies corresponds to the absence of radiation with other frequencies.

According to the quantum theory radiation is emitted on account of interchanges between states with energy levels E and E', the frequency being calculated from the principle of combination

$$E' - E = h\nu \qquad \ldots\ldots(6).$$

In the case just considered Planck's formula for energy levels is

$$E = nh\omega \qquad \ldots\ldots(7),$$

where n is integral.

Thus from equation (6) we obtain

$$\nu = (n' - n)\,\omega \qquad \ldots\ldots(8).$$

It appears from this that ν may occur in integral multiples of ω, but there is no indication of how intensely the corresponding spectral lines will appear. This is where the correspondence principle helps us. We refer to the equation (5) and note that only one frequency is represented. This corresponds to the value unity of $(n' - n)$ in equation (8), and since it is only in this case that the amplitude b of equation (5) has a value differing from zero, we conclude that only this frequency occurs.

Another illustration is obtained from a consideration of the case of the orbits of an electron moving under the attraction of a charged nucleus. According to the classical theory the radiation emitted would be in harmony with the fundamental frequency with which the electron traverses the orbit and with overtones if they are present.

Bohr's method of calculating the frequency of the electron in an orbit of order number n leads to the value $2R/n^3$ for the fundamental frequency, where $R = 2\pi^2 me^4/h^3$ is Rydberg's constant, m denoting the mass of the electron.

On the basis of the classical theory of the identity of radiation and mechanical frequencies we should expect that lines of frequency $2R/n^3\Delta n$ will be present in the spectrum, where Δn may have integral values.

The quantum theory, however, shows that the frequencies of the radiation are given by the formula

$$\nu = R\left(\frac{1}{n'^2} - \frac{1}{n^2}\right) \qquad \text{.....(9),}$$

where n and n' are integers.

This is quite a different result from that of the classical theory except for the case when n and n' are large numbers differing by integers much smaller than either. In this case we can write $n' = n + \Delta n$ and obtain the approximate form of equation (9):

$$\nu = \frac{2R}{n^3}\Delta n.$$

This has the form of the classical expression, and a correspondence between the classical and quantum frequencies appears for large quantum numbers. Bohr assumed that there exists also a correspondence between the intensities and states of polarization in the two cases, thus filling up the gap in the quantum theory by means of the classical theory. For, as we have seen, the quantum theory gives no information on these two characteristics of radiation, while the classical theory is precise on both. It is assumed that the information obtained in this way is exact for large quantum numbers and at least approximately correct for small quantum numbers.

The electric moment for a particular electron path may be described by the two components

$$P_x = \Sigma A_\tau \cos 2\pi\tau\omega t, \quad P_y = \Sigma B_\tau \sin 2\pi\tau\omega t \qquad \text{.....(10),}$$

where τ has the values 0, 1, 2, etc.

We obtain the following expressions for the electric intensity components by equation (1):

$$E_x \propto \ddot{P}_x = -\Sigma 4\pi^2\tau^2\omega^2 A_\tau \cos 2\pi\tau\omega t,$$

$$E_y \propto \ddot{P}_y = -\Sigma 4\pi^2\tau^2\omega^2 B_\tau \sin 2\pi\tau\omega t.$$

Thus corresponding to the frequency $\tau\omega$ the intensity of the radiation is proportional to $\tau^4(A_\tau^2 + B_\tau^2)$. In the classical theory τ denotes the order of the overtone, the fundamental frequency being ω.

It is understood in this theory that one path alone is concerned and that all the lines are emitted by a simultaneous process.

In the quantum theory, on the other hand, the emission of the different frequencies is by independent transitions between orbits, and this brings a difficulty into the procedure of the calculation. For there is no information as to whether equation (10) is to be worked out for the initial or for the final quantum orbit,

when we try to apply the correspondence principle. The riddle is solved in the new mechanics where both orbits or rather both numbers n and n', which refer to the orbits, occur symmetrically. The old quantum theory shed no light upon the difficulty.

Suppose that we make the representation for the final orbit, then the principle asserts that when the quantum numbers are small the classical process of calculation will give the approximate value of the intensities. The principle is to be regarded as a substitute for a statistical process which the old quantum theory failed to supply for itself. This theory was unable to give any information about the frequency of occurrence of the transitions responsible for a particular wave-length. It accepted the principle that the relative intensities could be obtained from the classical formula in the way we have illustrated by these examples.

References

(1) Bohr. Kopenhager Akademie (1918).
(2) Sommerfeld. *Atomic Structure and Spectral Lines.*

§ 3. THE WORK OF DE BROGLIE

The beginnings of the movement which have led to the foundation of that branch of the new quantum theory described as wave mechanics are to be found in some speculations of Louis de Broglie, who sought to associate the idea of periodicity with that of corpuscle. The latter was not to be thought of as a mere concentration of matter in a small region of space, but was to be associated with a "periodic phenomenon" extended throughout space. In order to give expression to the idea the restricted principle of relativity was applied. It would appear that this conception of the dual nature of a corpuscle is an example of remarkable intuition. It begins with the idea that the motion of a particle is inseparable from that of a wave. This is the basis of wave mechanics to which, as we shall see, Schroedinger's insight and mode of description have given a deeper significance.

de Broglie began by considering a system S' of coordinates which is in motion with respect to a system S with velocity v along a direction which serves as the x axis for both.

Thus if $(x'y'z')$ denote the coordinates of a point in S' and t' the time of an event at the point, the principle of relativity states that this point and event will be described in S by (xyz) and by t related to the S' values through the Lorentz transformations

$$x' = k(x - vt), \quad y' = y, \quad z' = z, \quad t' = k\left(t - \frac{v}{c^2}x\right) \qquad \ldots\ldots(11),$$

where $k = \left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}}$.

Consider a point at rest in S', and let it be associated with a stationary wave

$$Y' = f \sin 2\pi\nu' t' \qquad \ldots\ldots(12).$$

The function f can be any function of the coordinates $(x'y'z')$, but for the sake of simplicity let us take an example familiar in the subject of sound or light,

$$Y' = 2\cos\frac{2\pi\nu'}{c}x' \sin 2\pi\nu' t' \qquad \ldots\ldots(13).$$

This will be recognized as representing the superposition of two sine waves travelling in opposite directions, for example, along a stretched string. It is the simplest kind of wave motion that can be associated with a stationary particle.

The quantity Y' has a value for all values of x', and the wave may be said to occupy all space but the associated particle occupies a point of space.

In order to examine the association between a moving particle and a wave we have to see what Y' becomes from the point of view of the system S in which the particle is regarded as moving with velocity v.

From the Lorentz transformations we see that Y' becomes

$$Y = 2\cos\frac{2\pi\nu'}{c}k(x-vt)\sin 2\pi\nu' k\left(t-\frac{v}{c^2}x\right) \qquad \ldots\ldots(14).$$

The cosine part of this expression has a period of $c/k\nu' v$, while the sine factor has a period of $1/k\nu'$. The latter is very much shorter than the former, and we can look upon Y as representing a progressive wave with a relatively slowly varying amplitude of magnitude given by the cosine term. The wave travels forward with phase velocity $u = c^2/v$ and has the frequency $\nu = k\nu'$. We thus note the two relations

$$\nu = \nu_0/\sqrt{(1-v^2/c^2)}, \quad u = c^2/v, \qquad \ldots\ldots(15),$$

where we have written ν_0 instead of ν' to indicate that this frequency is associated with the particle at rest.

If we move along the axis of x with velocity v the expression $(x-vt)$ preserves the same value, so that we keep up with a constant amplitude of the wave. Thus the particle velocity is also the velocity of movement of a constant wave amplitude. Equation (15) gives the relation between this velocity and the phase velocity.

In optics we should describe the ratio c/u as the refractive index for the waves in a medium where their velocity is u. We obtain therefore, from equation (15), a dispersion formula.

For $\nu_0^2/\nu^2 = 1 - v^2/c^2 = 1-(c/u)^2$, whence

$$n = \sqrt{(1-\nu_0^2/\nu^2)} \qquad \ldots\ldots(16)$$

gives the relation between refractive index and frequency.

The velocity of the particle as a group velocity. We have just seen in the simple case of association of a wave with a particle that the velocity of the latter is the velocity of constant amplitude of a progressive wave with a slowly varying amplitude. But in order to obtain a deeper insight of the relation of the particle velocity to wave motion we must consider a group of waves. By a group of waves we mean the superposition of waves with frequencies varying over a small range. Denote $1/\lambda$ by N, the wave number, and let the corresponding frequency be denoted by ν_N. Suppose that a wave of this frequency be represented by

$$A_N \sin 2\pi(\nu_N t - Nx) \qquad \ldots\ldots(17).$$

A_N denotes the amplitude of the wave with the wave number N.

The group is represented by

$$\Sigma A_N \sin 2\pi (\nu_N t - Nx),$$

where the summation is made over a small range of values which lie about a mean frequency ν_0 and a mean wave number N_0.

In order to simplify the expression we shall replace the sines by the exponential functions and represent the group by

$$\Sigma A_N e^{2\pi i (\nu_N t - Nx)} \qquad \text{.....(18).}$$

We can take the real or imaginary part of this expression when one or the other is required. For our present purpose this separation is not necessary.

Let $\nu_N = \nu_0 + \nu'$ and $N = N_0 + N'$, where ν' and N' are small quantities.

The frequency depends upon the wave number so that we can write

$$\nu' = d\nu_N/dN \,.\, N'.$$

Thus equation (18) can be written in the form

$$e^{2\pi i (\nu_0 t - N_0 x)} \Sigma A_N e^{2\pi i \left(\frac{d\nu_N}{dN} t - x\right) N'},$$

where the summation extends over the small range of N'.

The quantity under the sign of summation can be treated as an amplitude which varies with x and t. As in the simple case which we first considered we have here a progressive wave with frequency ν_0 and phase velocity ν_0/N_0 with a fluctuating amplitude.

A_N does not depend upon x and t, so that if x and t vary in such a way that

$$\frac{d\nu_N}{dN} t - x = \text{a constant},$$

the amplitude remains constant. If we imagine an observer to travel with the wave with the velocity

$$\frac{dx}{dt} = \frac{d\nu_N}{dN},$$

he will always keep up with a constant amplitude of the superposed waves of the group. This velocity is the group velocity.

We can now omit the suffix and regarding ν as the mean velocity of the group we will write for the phase and group velocities

$$u = \nu/N \quad \text{and} \quad v = d\nu/dN \qquad \text{.....(19)}$$

respectively.

Since $n = c/u$, it follows that $n\nu = cN$ and

$$\frac{1}{c} \frac{d(n\nu)}{d\nu} = \frac{dN}{d\nu} = \frac{1}{v} \qquad \text{.....(20).}$$

Let us apply this to the case resulting from de Broglie's formula (16)

$$n\nu = (\nu^2 - \nu_0^2)^{\frac{1}{2}},$$

$$\frac{c}{v} = \frac{d(n\nu)}{d\nu} = \frac{\nu}{(\nu^2 - \nu_0^2)^{\frac{1}{2}}}.$$

From the first formula of equation (15) it can be seen that the v introduced in equation (19) as the group velocity is identical with the velocity of the particle. Thus the wave view of matter suggests the representation of the particle by a group of waves, the energy of the group supplying the mass of the particle by the relativistic identity of mass and energy which is expressed in the formula

$$W = mc^2 \qquad \text{.....(21).}$$

The relation between wave-length and momentum. The frequency we have introduced in the wave associated with the particle is assumed to be related to the energy of the particle by means of the formula

$$W = h\nu.$$

The wave-length λ can be expressed in terms of the momentum of the particle in the following way:

$$\lambda = \frac{u}{\nu} = \frac{c^2}{\nu v} = \frac{mc^2}{mv\nu} = \frac{W}{mv\nu} = \frac{h}{mv} \qquad \text{.....(22).}$$

Thus

$$\lambda = h/\text{momentum}.$$

This is de Broglie's law, and it has been shown experimentally in the cases of scattering of electrons and other particles that the process can be treated quantitatively as one of scattering of waves of this wave-length.

References

(3) Thèse de doctorat (Masson 1924).
(4) *J. Phys. Radium*, **1**, 1–6 (1926).
(5) *La Mécanique Ondulatoire*. Mémorial des Sciences Physiques, Fasc. 1 (1928). Gauthier-Villars et Cie.
(6) *Introduction to Wave Mechanics*, by de Broglie. Methuen and Co.

§ 4. MECHANICS AND OPTICS

The suggestions of de Broglie were a source of inspiration to Schroedinger, who formulated the hypothesis that the difficulties of atomic physics were of the character that the persistent and exclusive use of geometrical optics would encounter in the realm of physical optics. This is another example of a fine intuition which justifies the title of Schroedinger's Principle for this hypothesis.

In order to appreciate the significance of Schroedinger's work we must refer to the principles of classical dynamics.

It is an interesting fact that throughout the development of the quantum theory this part of the classical doctrine has played a very great part. The successes in astronomy were hardly greater than those of the few years in which it was applied to the minute solar systems of atoms. At the height of these successes it appeared as if dynamical methods had been waiting for this moment of greatest service to physics.

We shall see of what use it was in the development of wave mechanics, the foundation of which we owe to Schroedinger, and it was by an appeal to the mathematical forms employed in it that suggestions of the most fundamental character were obtained in matrix mechanics, to which we shall refer later.

We shall illustrate the general principles by means of a special case in order that the meaning of functions which appear in the theory may be appreciated by reference to a familiar example.

Let us consider the case of a particle of mass m which is projected upward with a velocity v_0 from a level which we shall choose as the origin of the coordinate x. Thus the x-axis lies in the vertical direction. The total energy is thus $W = \frac{1}{2}mv_0^2$.

For the purpose of the illustration it is necessary to draw attention to certain functions. One of these is the kinetic energy, T, and another the potential energy, V. In this case

$$T = \tfrac{1}{2}m\dot{x}^2, \quad V = mgx.$$

T is a homogeneous function of the second degree in the velocity $\dot{x}$. This is very often the form of T even when there are more velocity components than $\dot{x}$ and we shall limit ourselves to the cases where this is so.

Another important function is the Lagrange function

$$L = T - V.$$

In general cases L is a function of the velocities, coordinates and the time. The time may occur explicitly, but very often it is present only implicitly through the coordinates and velocity. We shall suppose that it does not occur explicitly.

In the present case L is a function of $\dot{x}$ and x.

The equation of motion

$$m\ddot{x} = -mg$$

can be expressed in terms of L by means of Lagrange's equation.

Note that

$$L = \tfrac{1}{2}m\dot{x}^2 - mgx,$$

$$\frac{\partial L}{\partial \dot{x}} = m\dot{x}, \quad \frac{\partial L}{\partial x} = mg.$$

Thus

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = 0 \qquad \text{.....(23)}$$

is identical with the equation of motion.

The momentum

$$p = m\dot{x} = \frac{\partial L}{\partial \dot{x}} \qquad \text{.....(24).}$$

The value of this method of describing the motion of any system does not appear from a simple case such as we have considered.

But the form of equation (23) and of equation (24) can be applied to very complicated cases. In these we have a number of coordinates which are usually described by $(q_1, q_2, \ldots, q_n)$, the number n depending upon the system considered. If the expression T for the kinetic energy be constructed, together with V for the potential energy, T will very frequently be a homogeneous quadratic in the $\dot{q}$'s and V some function of the q's, and t will not occur explicitly in many important cases to be considered. We then construct L as before, and for each variable q with its velocity $\dot{q}$ we write down the equation (23). The n equations so obtained are the equations of motion of the system.

We also obtain the expression, equation (24), for each coordinate

$$p_r = \frac{\partial L}{\partial \dot{q}_r} \qquad \text{.....(25),}$$

and we describe the p_r's as components of the generalized momentum, while the q_r's are the generalized coordinates.

Thus as the generalization of equation (23) we have Lagrange's equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_r}\right) - \frac{\partial L}{\partial q_r} = 0 \qquad \text{.....(26).}$$

Another function to be considered is

$$H = p\dot{x} - L \qquad \text{.....(27).}$$

We refer for the moment to our example,

$$H = p\dot{x} - \tfrac{1}{2}m\dot{x}^2 + mgx \qquad \text{.....(28),}$$

but H is not to be regarded as a function of p, $\dot{x}$ and x. A relation exists between p and $\dot{x}$, and it is always understood that $\dot{x}$ has been removed and H expressed as a function of p and x.

In the present case we must write, by means of $p = m\dot{x}$,

$$H = \frac{1}{2m}p^2 + mgx.$$

This applies to the general case. We have n equations like equation (26) together with the generalization of equation (25),

$$H = \Sigma p_r \dot{q}_r - L \qquad \text{.....(29),}$$

and by means of equations (25) and (29) H is able to be expressed as a function of the p's and q's. In general it may also contain t as an explicit variable, but we shall have no occasion to refer to such cases.

Let us turn to the special case of equation (28), and suppose that p, $\dot{x}$ and x undergo a variation.

Then $$dH = p\,d\dot{x} + \dot{x}\,dp - m\dot{x}\,d\dot{x} + mg\,dx.$$

The terms containing $d\dot{x}$ disappear, since $p = m\dot{x}$, thus

$$dH = \dot{x}\,dp = mg\,dx.$$

In this equation dH is expressed in terms of variations of p and of x of which H is a function. It follows from the equation that

$$\frac{\partial H}{\partial p} = \dot{x}, \quad \frac{\partial H}{\partial x} = mg = \frac{\partial V}{\partial x}.$$

In the particular example chosen mg is also equal to $-\partial L/\partial x$, and we can write

$$\frac{\partial H}{\partial x} = -\frac{\partial L}{\partial x}.$$

If the same procedure is applied to equation (29) these results are generalized, and we obtain

$$\frac{\partial H}{\partial p_r} = \dot{q}_r, \quad \frac{\partial H}{\partial q_r} = -\frac{\partial L}{\partial q_r} \qquad \text{.....(30).}$$

From Lagrange's equation (26), since $p_r = \dfrac{\partial L}{\partial \dot{q}_r}$,

$$\frac{dp_r}{dt} = \frac{\partial L}{\partial q_r} = -\frac{\partial H}{\partial q_r}.$$

The two sets of equations

$$\frac{dq_r}{dt} = \frac{\partial H}{\partial p_r}, \quad \frac{dp_r}{dt} = -\frac{\partial H}{\partial q_r} \qquad \text{.....(31)}$$

are known as Hamilton's equations.

If these be applied to our special example using the appropriate form for H, it will be seen that they lead to the equation of motion.

The equations (26) and (31) find their application in the description of the motion of particles of matter either singly or collected together in the form of complicated systems. It is in many cases possible to study a problem in mechanics completely by means of these equations.

But the problem may be approached by a different method which makes the solution depend upon that of a differential equation known as the Hamilton-Jacobi equation.

Let us consider the integral

$$S = \int_{t_0}^{t_1} L \, dt \qquad \text{.....(32).}$$

In the simple case we have been considering L is a function of $\dot{x}$ and x, and these are both functions of t.

Let the curve AB denote any functional relation between x and t, a particular case which will concern us is the case where the relation is that appropriate to the mechanical problem, e.g.

$$x = v_0 t - \tfrac{1}{2} g t^2.$$

The point A has coordinates (t_0, x_0) and B (t_1, x_1).

Let $A'B'$ denote a curve close to AB, and let A' and B' have coordinates $(t_0 + \Delta t_0, x_0 + \Delta x_0)$ and $(t_1 + \Delta t_1, x_1 + \Delta x_1)$ respectively.

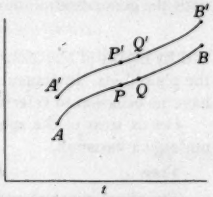


Figure 1

We pass from a point P on AB to P' on $A'B'$ with the same value of t by replacing x by $x + \Delta x$ and similarly $\dot{x}$ by $\dot{x} + \Delta \dot{x}$.

Thus the value of S for $A'B'$ can be written

$$S + \Delta S = \int_{t_0 + \Delta t_0}^{t_1 + \Delta t_1} (L + \Delta L) \, dt,$$

where

$$\Delta L = \frac{\partial L}{\partial \dot{x}} \Delta \dot{x} + \frac{\partial L}{\partial x} \Delta x.$$

The integral can be divided into three parts over the ranges

$$\int_{t_0}^{t_0} , \quad \int_{t_0}^{t_1} \text{ and } \int_{t_1}^{t_1 + \Delta t_1}.$$

The first range will contribute to the whole integral the quantity $-\Delta t_0\,(L+\Delta L)_0$, where $(L+\Delta L)_0$ means the value of this quantity close to (t_0, x_0). The variations are small and $\Delta L_0 \Delta t_0$ can be neglected in comparison with $L_0 \Delta t_0$ and L_0 means the value of L at the point A. Similar remarks show that the contribution in the third part of the range is $L_1 \Delta t_1$.

The second part contributes

$$\int_{t_0}^{t_1} L\,dt + \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial \dot{x}} \Delta \dot{x} + \frac{\partial L}{\partial x} \Delta x \right) dt.$$

It thus follows that the change in the value of the integral suffered in displacing the path of integration from AB to $A'B'$ is

$$\Delta S = (L_1 \Delta t_1 - L_0 \Delta t_0) + \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial \dot{x}} \Delta \dot{x} + \frac{\partial L}{\partial x} \Delta x \right) dt.$$

By making use of the relation $\Delta \dot{x} = \dfrac{d}{dt} \Delta x$ and integrating by parts the first term under the sign of integration, we obtain

$$\Delta S = \left[L \Delta t + \frac{\partial L}{\partial \dot{x}} \Delta x \right]_{t_0}^{t_1} + \int_{t_0}^{t_1} \left\{ \frac{\partial L}{\partial x} - \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{x}} \right) \right\} \Delta x \,.\, dt \qquad \text{.....(33).}$$

If the initial path corresponds to the motion of the particle, according to the laws of mechanics the expression within the brackets under the sign of integration vanishes (equation (23)).

If the system is complex and requires a number of variables for its complete description a similar argument leads to an expression for ΔS consisting of a summation of terms like those of the right side of equation (33) over all the variables, but the integral vanishes by equation (26).

We thus obtain the general result

$$\Delta S = \left[L \Delta t + \Sigma p_r \Delta q_r \right]_{t_0}^{t_1} \qquad \text{.....(34),}$$

where we write $p_r = \dfrac{\partial L}{\partial \dot{q}_r}$, equation (25).

Δq_r in equation (34) means a displacement with t fixed, but during the interval Δt P has become Q and the total displacement is QQ'. If we wish to express equation (34) in terms of the total displacement, δq, we can do so by means of the relation

$$\delta q = \Delta q + \dot{q} \Delta t \qquad (35).$$

If then we replace ΔS and Δt by δS and δt for the sake of uniformity of notation we can replace equation (34) by

$$\delta S = \left[(L - \Sigma p_r \dot{q}_r)\, \delta t + \Sigma p_r \delta q_r \right]_{t_0}^{t_1} \qquad \text{.....(36),}$$

$$= \left(-H \delta t + \Sigma p_r \delta q_r \right)_{t_0}^{t_1} \qquad \text{.....(37),}$$

by equation (29).

In the cases we are considering T is a homogeneous quadratic function of the $\dot{q}$'s and, as we saw in the example,

$$H=\frac{1}{2m}p^2+mgx=\text{total energy},$$

so in the general case $H=\Sigma p_r\dot{q}_r-L=\Sigma\dot{q}_r\frac{\partial L}{\partial\dot{q}_r}-L.$

But $\frac{\partial L}{\partial\dot{q}_r}=\frac{\partial T}{\partial\dot{q}_r}$, for $L=T-V$, and V does not contain $\dot{q}_r$.

Moreover, since T is homogeneous and of the second degree in $\dot{q}_r$, $\Sigma\dot{q}_r\frac{\partial T}{\partial\dot{q}_r}=2T.$

Thus $H=T+V.$

Let us suppose that the original and displaced path have the same end positions, i.e. δq at each end is zero, so that the second term on the right of equation (37) vanishes at each limit.

If H is constant during the motion, as is the case in the examples we consider, we can write

$$H=W,$$

e.g. in our example $H=\frac{1}{2}\frac{p^2}{m}+gx=\frac{1}{2}mv_0^2=W,$

where v_0 denotes the initial velocity at the level $x=0$.

In this case $\delta S=W(\delta t_1-\delta t_0)$ .....(38),

where δt_1 and δt_0 denote the time intervals at the ends of the displaced paths, and their difference denotes the difference between the times appropriate to the paths AB and $A'B'$. If we impose the further condition that these times shall be the same, we note that δS is zero or $S=\int_{t_0}^{t_1}L\,dt$ is stationary.

This is known as Hamilton's Principle.

We can deduce another similar principle from equation (37) by the introduction of the function

$$U=S+H_1t_1-H_0t_0=S+\left[Ht\right]_{t_0}^{t_1} \qquad .....(39).$$

We must note that we have gone back to the stage at which we had arrived in equation (37), removing the special conditions required to lead us to Hamilton's principle.

From equation (39), $\delta U=\delta S+\left[H\delta t+t\delta H\right]_{t_0}^{t_1}.$

Thus from equation (37)

$$\delta U=\left[\Sigma p_r\delta q_r+t\delta H\right]_{t_0}^{t_1} \qquad .....(40).$$

If, as before, we now impose the condition that the δq's vanish at the ends of the path, the first term on the right vanishes.

As before, H remains constant for the two paths but need not be the same constant for each. If we further impose the condition that the two paths are such that H is the same, i.e. $\delta H=0$, then $\delta U=0$. It is important to lay stress on these

stringent conditions, for it is only by means of them that the result (equation (37)) can lead to these two stationary principles, $\delta S = 0$ and $\delta U = 0$.

The latter, which is known as the principle of least action, but which should be described as the principle of stationary action, was discovered by Maupertuis. It can be stated very simply, for

$$U = S + \left[Ht\right]_{t_2}^{t_1}$$

$$= S + W(t_1 - t_2).$$

With our restriction $H = W$.

Thus
$$U = \int_{t_2}^{t_1} L\,dt + W\int_{t_2}^{t_1} dt$$

$$= \int_{t_2}^{t_1} (L+W)\,dt = \int_{t_2}^{t_1} 2T\,dt \qquad \text{.....(41).}$$

We have introduced these variational principles because of their importance in wave mechanics.

We have now to introduce the Hamilton-Jacobi differential equation by means of which particle dynamics may be given a new aspect.

Let us return to equation (37) and regard t_0 and the value of the coordinates at this time as constants, while we write $t = t_1$ and treat t and the coordinates at the upper limit of the path as current variables. We have

$$\delta S = -H\delta t + H_0\delta t_0 + \Sigma p_r \delta q_r - (\Sigma p_r \delta q_r)_0 \qquad \text{.....(42).}$$

Thus
$$\frac{\partial S}{\partial t} = -H, \quad \frac{\partial S}{\partial q_r} = p_r \qquad \text{.....(43).}$$

In general H depends upon the p's, q's and t which we can denote by writing

$$H = H(p_r, q_r, t).$$

Thus
$$\frac{\delta S}{\partial t} + H(p_r, q_r, t) = 0,$$

and from the second equation of (43) we obtain the following equation which is satisfied by S:

$$\frac{\partial S}{\partial t} + H\left(\frac{\partial S}{\partial q_r}, q_r, t\right) = 0 \qquad \text{.....(44).}$$

This is the Hamilton-Jacobi equation, and its importance is that if we can find a solution of the equation we can solve the mechanical problem. We do not require the proof of this statement for our present purpose, but merely wish to draw attention to the different methods of obtaining a solution, first, by solving the Lagrangian or the Hamiltonian equations (26) and (31), or secondly, by means of this differential equation (44). We shall limit ourselves to the simpler case when H does not contain the time explicitly.

Then equation (44) becomes

$$\frac{\partial S}{\partial t} + H\left(\frac{\partial S}{\partial q_r}, q_r\right) = 0 \qquad \text{.....(45),}$$

and since H is constant ($=W$) a solution is

$$S = -Wt + S_0(q_r) \qquad \ldots\ldots(46),$$

where S_0 is a function of the coordinates independent of t.

In order to make these points precise let us return to our particular example in which

$$H = \frac{1}{2m}p^2 + mgx = \tfrac{1}{2}mv_0^2 = W.$$

From equation (45) the equation is

$$\frac{\partial S}{\partial t} + \frac{1}{2m}\left(\frac{\partial S}{\partial x}\right)^2 + mgx = 0.$$

But from equation (46), which applies in this case since t is not present explicitly in H,

$$S = -Wt + S_0(x).$$

But

$$p = \frac{\partial S}{\partial x} = \frac{\partial S_0}{\partial x}$$

and

$$p = \sqrt{\{2m(W - mgx)\}} = m\sqrt{(v_0^2 - 2gx)}.$$

Thus

$$\frac{\partial S_0}{\partial x} = m\sqrt{(v_0^2 - 2gx)},$$

$$S_0 = -\frac{1}{3g}m(v_0^2 - 2gx)^{\frac{3}{2}} + \text{a constant}.$$

Thus

$$S = -Wt - \frac{m}{3g}(v_0^2 - 2gx)^{\frac{3}{2}} + \text{a constant} \qquad \ldots\ldots(47).$$

Schroedinger has given a very interesting geometrical significance to the function S and to some of the mechanical quantities associated with it. We shall first illustrate this by means of the example.

If we attribute a particular constant value to S in the case of equation (47) we obtain for a particular value of S at a particular time t,

$$-\frac{m}{3g}(v_0^2 - 2gx)^{\frac{3}{2}} = S + Wt = \text{constant} \qquad \ldots\ldots(48).$$

We can regard the constant as merged into the function S. In this example the space x of the system is one-dimensional, and if we give to S a particular value S_1 we can derive a series of values of x corresponding to a succession of values of the time. These can be represented in figure 2 by a series of lines drawn at different levels corresponding to the values of x.

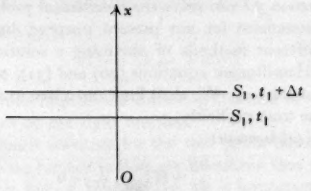


Figure 2

When the problem is one of two dimensions the locus for a particular value of S and of t is a curve, and when there are n coordinates $(q_1, q_2, \ldots, q_n)$ we can, by an extension of the geometrical language, consider the loci as n-dimensional surfaces.

Let us suppose that we follow the position of the series of values of x in figure 2, when S remains constant but t changes. The displacement between the values can be obtained from equation (48),

$$m(v_0^2 - 2gx)^{\frac{1}{2}}\Delta x = W\Delta t.$$

Thus if we travel along the x axis with the velocity given by $\Delta x/\Delta t$ we keep up with the same value of S.

The initial velocity is v_0, i.e. the velocity when $x=0$, so that we can, in the usual way, write

$$v_0^2 - 2gx = v^2,$$

v denoting the velocity at a distance x from the origin.

Thus we obtain
$$\frac{\Delta x}{\Delta t} = \frac{W}{mv} = \frac{W}{p} \qquad \ldots\ldots(49).$$

We can thus imagine the horizontal lines of the figure travelling with this velocity and having the property that they keep up with the constant value of S.

This cannot fail to remind us of the motion of a wave where we have a surface, the wave front, on which the phase remains constant as it moves. If we move so that we keep up with the wave front we keep up with a constant phase, and the velocity with which we move is the phase velocity.

A similar result can be obtained for the general case. We have

$$S = -Wt + S_0,$$

where S_0 is now a function of n coordinates q_r.

If S remains constant as we pass from surface to surface as the time progresses,

$$-W\Delta t + \Sigma \frac{\partial S_0}{\partial q_r}\Delta q_r = 0 \qquad \ldots\ldots(50).$$

But $(\partial S_0/\partial q_r)\,\Delta q_r$ denotes the scalar product of a vector with components $\partial S_0/\partial q_r$ and a vector with components Δq_r. We compare this with the simpler case of

$$\left(\frac{\partial F}{\partial x}\Delta x + \frac{\partial F}{\partial y}\Delta y + \frac{\partial F}{\partial z}\Delta z\right)$$

in which $(\partial F/\partial x, \partial F/\partial y, \partial F/\partial z)$ denote the components of a vector along the normal to the surface

$$F = \text{a constant}.$$

If $(\Delta x, \Delta y, \Delta z)$ are the components of the element between two neighbouring surfaces lying along the normal, Δn, the simpler expression can be written

$$(dF/dn)\,\Delta n \quad \text{or} \quad \text{grad}\, F.\Delta n.$$

Although the expression is more general we can write equation (50) in the form

$$-W\Delta t + \text{grad}\, S_0.\Delta n = 0.$$

Grad S_0 has the components $\partial S/\partial q_r$, etc., and its components are thus those of the momentum p_r, etc. If the momentum be described by p we have

$$\frac{\Delta n}{\Delta t}=\frac{W}{\text{grad } S_0}=\frac{W}{p} \qquad \text{.....(51).}$$

Thus we can describe W/p in the general case as the phase velocity

$$u=\frac{W}{p}=\frac{W}{\text{grad } S_0} \qquad \text{.....(52).}$$

We can thus interpret the quantity S as denoting something which travels in the space appropriate to the dynamical problem with the velocity equal to W/momentum.

The analogy between mechanics and optics. We have seen that the mechanical problem is usually solved by determining paths of particles. For example, in the simple case we have taken we found a solution in the form

$$x=v_0t-\tfrac{1}{2}gt^2.$$

In the general case of a system we have to complicate the geometry somewhat and find the orientation and position by means of n coordinates, but the solution is still obtained as a curve which represents the sequence of sets of values of the q's. We describe any such curve as a path of the system and it is perpendicular to the surfaces $S_0=$ constant.

The successive surfaces over which S_0 keeps the same value can be constructed by means of equation (51), for at any point we construct an element of the normal of length $W\Delta t/\text{grad } S_0$ and thus reach points where the value is the same at any instant Δt later. This corresponds to the construction of Huygens in optics for successive wave surfaces. We must note that the particle or point representing the system does not pass onward with the surface, for its velocity is not u but is proportional to $1/u$ (equation (49)).

Although we have mentioned the principle of Huygens, which in optics we associate with waves, we have not yet arrived at an analogy with physical optics. We are still concerned with rays.

There is a further striking analogy to be found by means of the principle of Maupertuis. Referring to equation (41) and the remarks which preceded it we saw that

$$\delta\int_{t_1}^{t_2}2T\,dt=0.$$

Since W is constant we can write

$$\delta\int_{t_1}^{t_2}\frac{2T}{W}\,dt=\delta\int_{t_1}^{t_2}\frac{p_r\dot{q}_r}{W}\,dt=\delta\int_{t_1}^{t_2}\frac{p_r\,dq_r}{W}.$$

But $p_r=\partial S_0/\partial q_r$, and thus $\quad p_r\,dq_r=\text{grad } S_0\,.\,dn=\dfrac{W}{u}\,dn \qquad \text{.....(53).}$

Thus $$\delta\int_{t_1}^{t_2}\frac{2T}{W}\,dt=\delta\int\frac{dn}{u} \qquad \text{.....(54).}$$

The element dn denotes an element of the path of the system and u is the phase velocity. Thus the right-hand side of equation (54) corresponds to Fermat's principle

in optics; a principle which is fundamental in the branch known as geometrical optics.

There is thus an analogy between geometrical optics and mechanics. This is by no means a new discovery, for it was known to Hamilton, who passed from optics through geometrical optics to mechanics. We owe to Schroedinger the procedure which takes us from mechanics through optics to wave mechanics. He has stressed the analogy between geometrical optics and ordinary mechanics, but he has had to make a conjecture in order to make more of the analogy than a trifle of analysis. In order to make the analogy fruitful it was necessary to recall that developments in optics in the branch known as physical optics, where wave-lengths are not treated as infinitely small in comparison with other lengths occurring in the problems, require us to leave rays and lay stress on waves. The difficulties which classical mechanics encountered in the treatment of atomic problems were conjectured to be of the same character as those encountered by geometrical optics in leaving the region of infinitely small wave-lengths. This procedure we have already described as Schroedinger's principle in recognition of the fact that he overcame the difficulties by leaving the equations of mechanics concerned with paths of particles and by beginning with a wave equation.

§ 5. THE WAVE EQUATION

There is very little before us to suggest the nature of the equation to which we are to refer instead of the equations of classical dynamics. The analogy with optics suggests that it will be a wave equation in which the coordinates of the system play the part of the space coordinates in the familiar case of light. The only definite thing we have is the phase velocity, equation (52), with the suggestion that the quantity S plays a part similar to that of phase.

Schroedinger made the suggestion, which has turned out to be a very fruitful one, that we should investigate a simple case by analogy with the case of the light wave.

In this case the wave equation is

$$\nabla^2 F - \frac{1}{u^2}\frac{\partial^2 F}{\partial t^2} = 0 \qquad \text{.....(55),}$$

where $\nabla^2 F$ means, as usual, $\Sigma \dfrac{\partial^2 F}{\partial x^2}$.

A commonly occurring case is that in which F has the form

$$F = A \cos 2\pi \left(\nu t - \frac{lx + my + nz}{\lambda}\right) \qquad \text{.....(56),}$$

where (l, m, n) denote the direction cosines of the direction in which the wave is travelling. A is a function of the coordinates only, so that t occurs in F in a very simple way.

In this case equation (55) can be written in the form

$$\nabla^2 F + \frac{4\pi^2\nu^2}{u^2} F = 0 \qquad \text{.....(57).}$$

There is nothing so far to suggest anything as to the nature of F, but the phase

$$2\pi\left(\nu t - \frac{lx+my+nz}{\lambda}\right)$$

corresponds to $$S = -Wt + S_0.$$

The phase is a dimensionless quantity, while S has the dimensions of action (i.e. energy × time). In order to obtain an expression for the phase in the wave of the extended mechanics, which we shall now describe as wave mechanics, we shall divide S by a constant of the dimensions of action and shall employ the expression $2\pi S/h$ for the phase.

The wave function will be described by the symbol ψ, and we shall suppose it to be of the simple construction

$$\psi = \psi_0 \cos\frac{2\pi S}{h}.$$

ψ_0 does not contain t, and the phase $2\pi S/h$ contains t in the expression $-2\pi Wt/h$. Thus the frequency ν is now given by

$$\nu = W/h$$

or $$W = h\nu \qquad \text{.....(58).}$$

This expression, which is the relation between energy and frequency introduced by Planck, is suggestive of a new significance for this relation. The wave equation thus becomes

$$\nabla^2\psi + \frac{4\pi^2}{h^2}\frac{W^2}{u^2}\psi = 0 \qquad \text{.....(59).}$$

We cannot be said to have derived this equation, for there is nothing in mechanics to lead us logically to it. That it should occupy a fundamental place in wave mechanics is the suggestion we derive from Schroedinger's principle; it is in fact the quantitative expression of the principle. Wave mechanics begins with it in the same way that ordinary mechanics begins with Newton's laws of motion, or, what is the same thing, with Hamilton's or Lagrange's equations or with the Hamilton-Jacobi equation.

To take a concrete example, let us return to the motion of the particle under gravity. If we wish to apply the wave equation in such a case we note from equation (49) that $u = W/mv$ and $\nabla^2\psi$ is now simply $\partial^2\psi/\partial x^2$.

Thus the equation is $$\frac{\partial^2\psi}{\partial x^2} + \frac{4\pi^2}{h^2}m^2v^2\psi = 0 \qquad \text{.....(60).}$$

We can write $$W = T + V = \tfrac{1}{2}mv^2 + V$$

or $$m^2v^2 = 2m(W - V),$$

and equation (60) becomes

$$\frac{\partial^2\psi}{\partial x^2} + \frac{8\pi^2 m}{h^2}(W - V)\psi = 0 \qquad \text{.....(61).}$$

ψ is a function of x, the space coordinate, together with the time t, and we might feel tempted to suppose that ψ has a physical counterpart, as also the waves themselves, such as we commonly attribute to light waves. The study of this question

leads us to think that these waves have as much right to be associated with a physical counterpart as waves of light, but the result of deliberations on this point has been to modify our views in the latter case. We shall return to this question later. But in the early tendency to allow ourselves to be influenced in this direction we have met with a difficulty, for we have seen that our simple example is only a special case of the general dynamics of systems and the space is then that of the coordinates, q_n. The wave equation belongs to this space as much as to what we should describe as real space. We must then apply equation (59) in the sense of this more general geometry and substitute u from equation (52), but the result is the same, we have again the equation (60) in the form

$$\nabla^2\psi + \frac{4\pi^2}{h^2} p^2\psi = 0.$$

We must remember that p^2 means $|\text{grad } S_0|^2$, and without going further into detail shall content ourselves with the statement that this expression can be shown to be equal to $2m(W-V)$, so that in the general case the wave equation is

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}(W-V)\psi = 0. \qquad \text{.....(62).}$$

We made a further demand of the function ψ which is that it should be unique and continuous over the whole of the region where it applies. This appears to be the least that can be asked of the function, but it is sufficient to replace the integral conditions

$$\int p\,dq = nh \qquad \text{.....(63)}$$

of the old theory.

When the wave equation is applied in a particular case it is found that only certain values of W are possible if this condition is to be satisfied. These are the characteristic values of the equation, and associated with them there are certain functions for ψ which are known as the characteristic functions.

The values of W are those of energy levels, e.g. in the simple treatment of the hydrogen atom the characteristic values of W are of the form

$$W_n = -\frac{2\pi^2 me^4}{n^2h^2} \qquad \text{.....(64),}$$

where n has integral values.

These are, of course, identical with the Bohr energy levels.

References

(7) *Ann. Phys., Lpz.*, **79**, **80**, **81** (1926).
(8) SCHROEDINGER, E. *Collected Papers on Wave Mechanics.* Blackie and Son.
(9) WHITTAKER, E. T. *Analytical Dynamics.* Camb. Univ. Press.
(10) WILSON, W. *Theoretical Physics*, **1**. Methuen and Co.
(11) In a form particularly applicable to the Quantum Theory. Notes at the end of Sommerfeld's *Atomic Structure and Spectral Lines.*
(12) BORN, M. *The Mechanics of the Atom.* G. Bell and Sons.

§ 6. QUANTUM OR MATRIX MECHANICS

While the work of de Broglie and of Schroedinger was developing, research in an apparently quite different direction was leading Heisenberg to a remodelling of mechanical equations by means of a notation peculiarly adapted to express the relations which had been found to exist between atomic quantities. He appears to have been influenced by the experience through which physicists had gone a quarter of a century earlier. At that time the ether dilemma had absorbed much energy and attention, but in looking back to this period one sees that the difficulties were chiefly those that scientists had themselves introduced. Einstein found a way out by doing without the ether and its deformities.

In reading Heisenberg's first contribution to what is described as quantum mechanics we see an attempt to throw overboard accumulated trappings in the hope that the freedom gained would do more than make up for the support lost.

He saw an objection to the formal rules of the existing quantum theory in that in the calculation of observable quantities it was necessary to construct rules which contained relations between fundamentally unobservable quantities as essential features. To illustrate such quantities we have only to call to mind positions in electron paths within the atom and periods of describing the paths. It might be hoped that such quantities are not fundamentally unobservable but are only so because of the inability of observers. If this is the nature of the difficulty in the observation then it would appear that a notation which makes use of such quantities cannot lead us far astray. But an examination of the success attending the use of such quantities has been very limited in the old quantum theory. Success in connexion with the hydrogen atom and in the Stark effect in this atom might give some ground for hope that the mode of description was suitable and promising of more success. But even in this atom disappointment is found in problems in which electric and magnetic fields occur in different directions, and as we have seen the failures are more impressive than the successes in the later developments where several electrons are concerned.

The purpose guiding the new step was to discover a new mechanics which retained relations between observable quantities omitting those between unobservable quantities. Heisenberg based the structure of the new mechanics on the frequency condition with the associated combination principle and upon Kramer's dispersion theory.

The frequency condition is

$$h\nu_{mn} = E_m - E_n \qquad \text{.....(65),}$$

and the combination principle is

$$\nu_{ln} = \nu_{lm} + \nu_{mn} \qquad \text{.....(66).}$$

This means that if two frequencies ν_{lm} and ν_{mn} occur in a spectral series it will be possible to discover a frequency ν_{ln} given by the formula.

There are exceptions to this rule which are provided for by certain limitations to it. This should be contrasted with what the classical theory of vibrating systesm leads us to expect. If we consider the example of a stretched string in vibration,

the frequencies expected consist of a fundamental ν_0 and its overtones $n\nu_0$, where n is an integer.

We can thus write as the classical principle of combination

$$l\nu_0 + m\nu_0 = (l+m)\,\nu_0 \qquad \text{.....(67).}$$

Let us fix our attention on a particular quantum state to which transitions occur. This state is characterized by an integer n; the corresponding energy, for example, is denoted by E_n. If we return to the conception of electron paths, and as in equation (10) describe the path in the classical manner, we shall find a series of possible frequencies, all integral multiples of ω, for the radiation emitted. This frequency ω depends upon n and we can denote this by adding a suffix to ω.

The other frequencies will depend upon other integers, one of which we denote by m, and a useful notation $\nu\,(n, m)$ will denote this dependence

$$2\pi\nu\,(n, m) = m\omega_n.$$

The rule (equation (67)) will be written

$$\nu\,(n, l) + \nu\,(n, m) = \nu\,(n, l+m) \qquad \text{.....(68).}$$

In the quantum theory the frequency ν_{ln} depends on two paths, and these are represented by the integers n and l. If we use a similar notation equation (66) becomes

$$\nu\,(l, m) + \nu\,(m, n) = \nu\,(l, n) \qquad \text{.....(69).}$$

We must note the difference between this and equation (68).

In discussing the correspondence principle we saw the importance of expressing a displacement x by means of a Fourier series. We shall use an exponential representation for our present purpose in the form

$$x\,(t) = \sum_m A_m\,(n)\,e^{i\omega_n m t} \qquad \text{.....(70).}$$

The presence of n denotes that this expansion is made for a particular quantum path characterized by n, and the summation is made over values of m. The meaning of ω_n is, as we have just described, the fundamental frequency of the path characterized by n. x is made up of terms containing certain amplitudes $A_m(n)$, frequencies $m\omega_n$ and phases. We can look upon x as describing the path or the particle in the path without laying too much stress upon it as a coordinate for the position of the particle in the orbit. The expression, equation (70), has the advantage that it contains the observable quantities, amplitude, frequencies and phases which are capable of direct observation in the radiation, and it is a representation exactly in the character of Heisenberg's proposal. But we have seen that it has the disadvantage that it lays stress on what we should describe in the old quantum theory as one orbit and does not mention the other associated with it, which that theory requires for a transition to produce radiation. We shall now avoid using the term orbit and speak of state, and we can express this disadvantage by saying that the formula gives too special a significance to n. We know that in the quantum theory the two numbers n and m occur symmetrically.

This is not the case with n and m in equation (70), and their meanings in regard to frequencies are quite different.

Considerations of this character led Heisenberg to modify equation (70) so that the two numbers n and m appeared symmetrically in the formula for x.

If we turn to the notation of equations (68) and (69) we see that the modification suggested is

$$x(t)=\sum_{m} A(n,m)\, e^{2\pi i\nu(n,m)\,t} \qquad \ldots\ldots(71).$$

In order to appreciate Heisenberg's next step we must return to the Fourier expansion (equation (70)) and examine the process of multiplication of two such quantities as x and y.

Let us suppose that y is expanded in the same way as x, but with coefficients $B_l(n)$ as characteristic of it:

$$y(t)=\sum B_l(n)\, e^{i\omega_n l t}.$$

The summation is over l which we use in the place of m to avoid confusion:

$$xy=\sum_{m}\sum_{l} A_m(n)\, B_l(n)\, e^{i\omega_n(l+m)\,t} \qquad \ldots\ldots(72).$$

By writing $(l+m)=k$ we can replace this by

$$\begin{aligned} xy&=\sum_{m}\sum_{k} A_m(n)\, B_{k-m}(n)\, e^{i\omega_n k t} \\ &=\sum_{k} C_k(n)\, e^{i\omega_n k t} \qquad \ldots\ldots(73). \end{aligned}$$

This product is now in the same form as the expressions for x and y with

$$C_k=\sum_{m} A_m(n)\, B_{k-m}(n).$$

The rule for combination of the indices of the exponential is that of simple addition,

$$\omega_n m+\omega_n l=\omega_n k,$$

and is identical with the classical law of combination of frequencies, equation (67) or (68).

Thus the ordinary analysis of algebra provides that this principle of combination shall be incorporated in the notation used in the classical treatment of vibrations.

Let us see how we can provide for the principle, equation (69), in a notation which might be suitable for quantum mechanics.

This formula will be included in such a product of x and y provided that $e^{2\pi i\nu(n,m)\,t}$ and $e^{2\pi i\nu(m,l)\,t}$ come together in a product term, for we could then obtain $e^{2\pi i\nu(n,l)\,t}$ of the law (equation (69)).

We consider now the expression (71) and another like it:

$$y(t)=\sum_{p} B(p,l)\, e^{2\pi i\nu(p,l)\,t} \qquad \ldots\ldots(74).$$

It will be noted that we have changed the positions of the indices compared with those of equation (71). The reason will appear from a consideration of the nature of the analysis immediately below.

The product

$$xy=\sum_{m}\sum_{p} A(n,m)\, B(p,l)\, e^{2\pi i\{\nu(n,m)+\nu(p,l)\}\,t}$$

and we enquire if the product can be reduced to the form of x or y.

That is to say, can we determine a form of analysis which makes

$$xy = \sum_l C(n, l)\, e^{2\pi i \nu (n, l)\, t}?$$

This will be the case if the product consists of a sum of terms in which only equal values of m and p occur together.

We shall then have in the index

$$\nu(n, m) + \nu(m, l),$$

which, according to the quantum principle of combination, gives $\nu(n, l)$ for all values of m.

This last statement follows at once from Bohr's relation between frequency and energy levels,

$$\nu(n, m) = \frac{E_n - E_m}{h}.$$

Thus if $$C(n, l) = \sum_m A(n, m)\, B(m, l)$$

the form of xy is that of the individual factors.

This form of product is recognized as a form of multiplication occurring in vector analysis or matrix analysis and also in the multiplication of determinants.

It is interesting that Heisenberg does not appear to have been familiar with this analysis and was obliged to develop this notation of multiplication in order to express his ideas. This makes his discovery all the more remarkable, but immediately after his discovery he was assisted by Born and Jordan, who recognized the nature of the analysis.

Having recognized the nature of the operations we need not labour its discovery any further but can proceed at once to the matrix operations which give the name "Matrix Mechanics" to this aspect of quantum mechanics.

The first thing to consider is the meaning of the term "matrix". We can proceed to an understanding of what a matrix is through vector analysis, which is well known to physicists.

Let us consider a line OA to represent a vector drawn from a convenient origin O. The vector is represented by the line OA but is often described in terms of its components (x, y, z) which are the coordinates of the point A referred to a set of axes.

We can write $$OA = (x, y, z) \qquad \text{.....(75),}$$

if we wish to lay stress on the components.

The unit vectors along the axes may be denoted by $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, and in the notation of vectors we write

$$\overrightarrow{OA} = \mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \qquad \text{.....(76).}$$

We pass over the familiar scalar and vector products to an operator known as a dyadic. This is written in the form

$$\mathbf{a} = a_{11}\mathbf{ii} + a_{22}\mathbf{jj} + a_{33}\mathbf{kk} + a_{12}\mathbf{ij} + a_{21}\mathbf{ji} + \text{etc.} \qquad \text{.....(77),}$$

or we may drop the vectors as in equation (75) and write

$$\mathbf{a} = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{Vmatrix} \equiv \| a_{mn} \| \qquad \text{.....(78).}$$

Of course the expression (78) is not a determinant, and the double vertical lines are drawn to call attention to that fact. It corresponds to the expression of a vector in terms of its components by equation (75). The dyadic $\mathbf{a}$ is to be regarded as an entity, the quantities a_{mn} as components. It is written in the matrix form in equation (78), and when expressed in this way $\mathbf{a}$ is usually called a matrix. We have taken an example of a matrix with three rows and three columns, but the number is not limited to three; it may have any value and may be infinite, as is the case in many of the matrices occurring in the quantum theory.

We can write the expression (77) in the form $a_{mn}\mathbf{i}_m\mathbf{i}_n$, assuming that summation occurs for the suffixes which are alike and that $\mathbf{j}=\mathbf{i}_2$, $\mathbf{k}=\mathbf{i}_3$. We have spoken of this quantity as an operator. This is because, unlike the scalar or vector product of two vectors, it contains no sign of operation between its vector parts $\mathbf{i}_m$ and $\mathbf{i}_n$, e.g. $\mathbf{i}_m.\mathbf{i}_n$ or $\mathbf{i}_m\times\mathbf{i}_n$. It is awaiting a quantity upon which it can operate. Suppose we take a vector $\mathbf{A}=A_1\mathbf{i}_1+A_2\mathbf{i}_2+A_3\mathbf{i}_3=A_l\mathbf{i}_l$ again assuming summation over like suffixes. We can now allow $\mathbf{a}$ to operate scalarly upon $\mathbf{A}$ forming $\mathbf{a}.\mathbf{A}$. Remembering that scalar multiplication between unlike vectors $\mathbf{i}$, $\mathbf{j}$ or $\mathbf{k}$ gives zero and that $\mathbf{i}.\mathbf{i}=\mathbf{i}$, we have $\mathbf{i}_m.\mathbf{i}_n=\delta_{mn}$, where δ_{mn} is unity if m and n are equal and zero if unequal. Thus $\mathbf{a}.\mathbf{A}=a_{mn}A_n\mathbf{i}_m$, and the operation has resulted in a vector with components $(a_{mn}A_n)$, where m has the values 1, 2 or 3, and in each case summation is implied over n. But we must also consider the result of multiplying two matrix operators together. This means the result of writing down a matrix $a_{mn}\mathbf{i}_m\mathbf{i}_n$ and $b_{pq}\mathbf{i}_p\mathbf{i}_q$ and multiplying $\mathbf{i}_n$ and $\mathbf{i}_p$ scalarly. The suffixes p, q are written in the matrix $\mathbf{b}$ in order to avoid confusion with m and n. In both cases the suffixes are dummies for summation is implied.

Thus the product is

$$\mathbf{a}.\mathbf{b} = a_{mn}b_{pq}\mathbf{i}_m\mathbf{i}_n.\mathbf{i}_p\mathbf{i}_q = a_{mn}b_{pq}\delta_{np}\mathbf{i}_m\mathbf{i}_q$$
$$= a_{mn}b_{nq}\mathbf{i}_m\mathbf{i}_q.$$

We note that the operation has resulted in the formation of another matrix of the form $c_{mq}\mathbf{i}_m\mathbf{i}_q$, where $c_{mq}=a_{mn}b_{nq}$.

In general $\mathbf{b}.\mathbf{a}$ is not equal to $\mathbf{a}.\mathbf{b}$. In the cases when this is so $\mathbf{a}$ and $\mathbf{b}$ are said to be commutative, but we must regard this as a special case, for it is a commonplace of matrix algebra that multiplication as here defined is non-commutative. If we turn back to the expression $C(n, l)$ we see that it can be regarded as the component of a matrix $\mathbf{C}$ which results from the product of matrices $\mathbf{A}$ and $\mathbf{B}$.

The product xy which we considered in connexion with $C(n, l)$ can now be recognized as part of a matrix product. It is the nl component of the product of two matrices which we can describe by $\mathbf{x}$ and $\mathbf{y}$. We write

$$\mathbf{x} = \| A_{nm}e^{2\pi i\nu_{nm}t} \|, \quad \mathbf{y} = \| B_{pl}e^{2\pi i\nu_{pl}t} \|,$$

and $\mathbf{x}.\mathbf{y}$ is exactly of the same character as $\mathbf{a}.\mathbf{b}$ which we have just considered.

Thus the x which Heisenberg expanded in the form of a modified Fourier's series, equation (71), is to be still further generalized and to be recognized as a matrix. In this generalization the difficulty inherent in the correspondence principle that the quantum states did not occur symmetrically disappears at once. The suffixes n and m occurring in the components of x refer to the states and A_{nm} and ν_{nm} measure quantities which we describe as due to transitions between those states. We can look upon this identification of x with a matrix as the expression of the correspondence principle in its complete form.

It is important to realize how far we have gone from the old conception of the atom in thus introducing a matrix to describe the radiation emitted by it. It is true that we began with an analogy in which a quantity x was regarded as locating an electron, but the idea that location of the electron is possible is not included in the form of description introduced by Heisenberg. He began with the idea of elimination of orbits and location in them, and the matrix notation is just what he required to carry out his object. The components of $\mathbf{x}$, e.g. $A_{nm}e^{2\pi i\nu_{nm}t}$, represent the amplitude, phase and frequency of the radiation associated with the transition from the state n to the state m. Thus A_{nm} is itself a complex quantity and can be written in the form $a_{nm}e^{i\beta_{nm}t}$, where a_{nm} is real and represents the amplitude, β_{nm} represents the phase and ν_{nm} is the frequency, which is given by means of

$$\nu_{nm}=\frac{E_n-E_m}{h},$$

where E_n and E_m denote energies associated with the respective states.

The general use of matrices in the mechanics of the atom. The case we have been considering suggests that the coordinates of position must be replaced by matrix quantities, and it seems natural to extend this replacement. The conclusion is that the quantum theory of the atom is to be described by means of the notation of matrices. In order to carry out the plan that is thus placed before us we have to represent the quantities between which relations exist by means of matrices and to discover equations connecting them. These equations are to represent the laws of the new mechanics. In the attempt to find relations that might reasonably be expected to lead to fruitful results we have no guide except that which classical mechanics and the old form of the quantum theory provide. We have already made use of the comparison in replacing the familiar Fourier expansion by means of another which bears an outward resemblance to it. Thus the first suggestion is to replace all familiar quantities in mechanics by means of matrices and to write down equations relating them which have the outward form of the equations of ordinary mechanics and to impose limitations suggested by the form of the old quantum theory. In this foundation of the subject there is necessarily prolonged research, much of it has already served its purpose and now belongs to the early history of the subject. We recognize that the equations we use are in the last resort the basic assumptions of the theory. We accept them as we accept the law, $P=mf$, as the basis of Newtonian dynamics. One of the most interesting of the new relations is

$$\mathbf{pq}-\mathbf{qp}=\frac{h}{2\pi i}\mathbf{I} \qquad \text{.....(79).}$$

This is a quantum condition corresponding to $\int p dq = nh$ of the older theory.

$\mathbf{p}$ is the momentum matrix and $\mathbf{q}$ the coordinate matrix and they have components

$$p_{mn} e^{2\pi i \nu_{mn} t} \quad \text{and} \quad q_{mn} e^{2\pi i \nu_{mn} t}.$$

Both p_{mn} and q_{mn} are complex and we can simplify the notation by understanding that the exponential factors occur with these quantities. The right-hand side contains Planck's constant and $\mathbf{I}$ is understood to mean the unit matrix, i.e. one which only contains diagonal elements I_{11}, I_{22}, etc., each of which has the value unity. All such elements as I_{12}, I_{23}, are zero.

Thus equation (79) means

$$\sum_n (p_{1n} q_{n1} - q_{1n} p_{n1}) = \frac{h}{2\pi i},$$

together with similar relations in which 1 is replaced by 2, 3, etc. This relation cannot be satisfied unless $\mathbf{p}$ and $\mathbf{q}$ are matrices with an infinite number of rows and columns.

The object of the analysis is to find values for the components of $\mathbf{q}$ and $\mathbf{p}$ and to find values of the energy levels and of the frequencies. Heisenberg and others have developed the analysis so that these calculations can be made, and have in this way obtained results which verify that the notation is more suitable for the description of atomic phenomena than that used previously. His earliest attempt to solve the problem of the harmonic oscillator, which has always played a fundamental part in quantum theories, led to energy values $(n + \frac{1}{2})\, h\nu$, so that the half-quantum number came in quite naturally and was not forced into the theory empirically by the requirements of experiment.

The methods introduced by wave mechanics and matrix mechanics appear to be widely different from one another. As the methods have been presented here it may appear that the latter is the more complete. It is a self-contained analysis and it gives us the energy values and the amplitudes of the radiations. Wave mechanics, although it gives the energy values, appears to say nothing on the question of intensities. But Schroedinger was able to show how intensities could be calculated in accordance with the principles he had introduced into wave mechanics, and he showed very early in the development of his work the relation of the two methods to one another. The result has been that the apparently rival methods have been of great assistance to each other. Wave mechanics considers the problems of experiment in an intuitive form and makes use of a highly developed mathematical analysis by processes which are classical in character, and it is thus possible to get quickly under way in our attempts at expression. Matrix mechanics is a less explored analysis, and to those trained in classical methods appears less directly related to the objects of its description, but it has been greatly helped as a result of Schroedinger's discovery of its connexion with wave mechanics. Dirac has shown that quantities which he calls q numbers more general than matrices satisfy the laws of motion and conditions of quantum mechanics. His work unites the two procedures we have outlined and goes beyond them.

§ 7. THE UNCERTAINTY PRINCIPLE

If the principle described here as Schroedinger's principle be accepted, namely, that we have to study the behaviour of particles by means of the wave equation, we can derive a very striking consequence which was developed in collaboration by Bohr and Heisenberg. The theory is based especially upon the researches of the late Lord Rayleigh on the resolving power of optical instruments. We shall derive the necessary relations by the consideration of familiar examples.

Let us consider an example from the theory of sound. If two sources of sound are vibrating together and emitting waves of frequencies ν and $\nu+\Delta\nu$, beats are produced of frequency $\Delta\nu$. In order to appreciate that the frequencies are different we shall require to hear at least one beat, that is to say, we shall require an interval of time Δt, where Δt is of the order of magnitude $1/\Delta\nu$. Thus between the interval of time required to measure the frequency difference and the difference itself there exists the relation

$$\Delta\nu\Delta t\sim 1 \qquad \text{.....(80).}$$

The symbol $\sim$ is to be understood to mean that the quantities on either side of it are of the same order of magnitude. There is nothing in this relation which is peculiar to sound, the same is true of any wave theory provided that we accept the principle of superposition. But to continue with this example, suppose that a series of sources of sound with frequencies lying within the range ν to $\nu+\Delta\nu$ is examined, we shall require an interval of time Δt in order to detect this range of frequency. We can say that in time Δt no frequency difference can be measured which has a value less than a quantity of the order $1/\Delta t$.

We can obtain another relation of a similar character from the familiar example of the resolving power of a diffraction grating. This is usually written in the form

$$\frac{\lambda}{\Delta\lambda}=Nn \qquad \text{.....(81),}$$

where N is the number of elements in the grating and n the order of the spectrum. The usual mode of derivation of this formula recalls that its meaning is that if two trains of waves of wave-lengths λ and $\lambda+\Delta\lambda$ fall upon a grating, they will be distinguishable as waves of different lengths only if $\Delta\lambda>\lambda/Nn$. Thus if a group of waves of a range of wave-lengths λ to $\lambda+\Delta\lambda$ fall on the grating, it will be impossible to measure $\Delta\lambda$ unless this relation holds.

In figure 3 let AB denote the grating and suppose that the waves fall normally upon it. Let AC be the direction in which the spectrum of the order n is seen, so that $AC=Nn\lambda$, where C is the foot of the perpendicular from B on to AC.

Thus the relation for resolving power means that we require at least the length AC to distinguish $\Delta\lambda$. According to Rayleigh's criterion for resolution this means that the length AC is necessary in order that the central fringe due to one of the waves may lie over the first dark fringe due to the other. Thus the wave train containing this range of frequencies must be at least as long as $AC\,(=Nn\lambda)$ in order that it may be engaged by the resolving instrument in the process of resolution.

Let this length of the wave train be denoted by Δs. The relation between Δs and $\Delta\lambda$ which must exist in order that $\Delta\lambda$ may be distinguished is

$$\Delta\lambda > \frac{\lambda^2}{\Delta s}.$$

We can thus say that for resolution

$$\Delta s \Delta\left(\frac{1}{\lambda}\right) \sim 1 \qquad \ldots\ldots(82).$$

We can thus say that in a determination of the frequency ν we cannot distinguish between ν and $\nu+\Delta\nu$ unless the time at our disposal, Δt, satisfies equation (81), and in the determination of the wave number $1/\lambda$ we cannot distinguish between it and $1/\lambda+\Delta(1/\lambda)$ unless the length of wave at our disposal satisfies equation (82).

These relations are quite general. The examples we have chosen help to make their meaning precise, but they impose no special limit to their application. They show that the introduction of groups of waves is associated with an uncertainty with

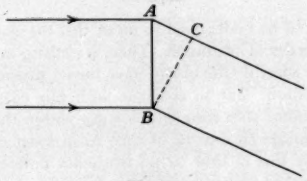


Figure 3

regard to their frequency range and their range of wave number or of wave-length. The groups are limited in their extent, and this limitation is brought about by the principle of superposition.

In order to obtain an idea of how this is brought about we first note that a plane monochromatic wave represented by

$$Y = A \sin 2\pi(\nu t - r/\lambda) \qquad \ldots\ldots(83),$$

where r denotes a length along the direction in which the wave is travelling, has a value of Y for all values of the time t and of the length r. It may be said to occupy all time and space and is thus of the nature of a mathematical abstraction, since the waves of physics occupy a limited range both in space and time.

The latter are wave trains and are represented as a sum of such expressions as (83) with different values of A and of ν and with different phases. In addition the amplitudes and phases may depend upon the direction of travel of the waves making up the train. A collection of such plane waves with a small range of frequencies and with neighbouring directions forms a wave group.

Under certain circumstances this group can be represented, as we have already seen, by a function F of the form

$$F = C \sin 2\pi(\nu t - r/\lambda) \qquad \ldots\ldots(84).$$

The amplitude C depends upon the coordinates and time and ν and λ are the mean frequency and wave-length of the members of the group.

If such an expression can represent a group, the value of C at any particular time will be zero except over a limited region of space, and at any point it will be zero except over a limited region of time. These conditions imply severe limitations upon the quantities occurring in equation (84). Suppose that the direction cosines of r are (l, m, n), then the quantities l/λ, m/λ, n/λ may be regarded as the components of the wave number $1/\lambda$ in the directions of x, y and z. The wave number thus appears to be a vector, a fact which agrees well with the de Broglie representation (equation (22))

$$\text{momentum} = h/\lambda.$$

The limitations which have to be satisfied in order to make it possible to represent a group in this way are

$$\delta(l/\lambda)\,\delta x \sim 1, \quad \delta(m/\lambda)\,\delta y \sim 1, \quad \delta(n/\lambda)\,\delta z \sim 1$$
$$\delta\nu\,\delta t \sim 1 \qquad \text{.....(85).}$$

The mathematical treatment of the question leads to the result that these products cannot be of an order less than unity.

We have seen that energy and momentum are related with a frequency and wave-length by the formulae $W = h\nu$, equation (58), and $p = h/\lambda$, equation (22). These formulae contain sharp contrasts, especially when we apply them to a particle, for W and p belong to the particle conception of matter while ν and λ belong to the conception of a monochromatic wave and are associated with unbounded space and time. In order to make use of the wave conception in the representation of a particle we have to bring in the principle of superposition and form a group in the way we have described.

As we have already seen (equation (20)), the group velocity is to be identified with the velocity of the particle.

With the identification of mechanical with wave quantities which equations (22) permit we obtain from (85) the following uncertainty relations:

$$\delta p_x \delta x \sim h, \quad \delta p_y \delta y \sim h, \quad \delta p_z \delta z \sim h, \quad \delta W \delta t \sim h \qquad \text{.....(86).}$$

Just as equation (85) describes the limits to which measurements of wave-lengths and frequencies by optical instruments are subjected, and just as the representation of a group by a wave train is possible only under these conditions, equation (85), so equation (86) shows the limits of accuracy of energy and impulse measurements and the limits to their definition.

Suppose we represent a particle by means of a group of waves of short extent. The position of the particle may be described by locating it at some point within the group. But we cannot say that it lies at one point more than at another. Heisenberg's interpretation of relations (86) is that the relations indicate the greatest possible accuracy with which space-time coordinates and the values of momentum and energy can be measured. If in any circumstances it is possible to locate a particle with an accuracy Δx, then according to these relations the momentum can be determined only to the accuracy Δp, where $\Delta p \Delta x \sim h$. It is only with this

limitation that physical phenomena can be described by means of particles. The success of this mode of description hitherto in large-scale theories in physics is due to the fact that h is very small.

It will be remembered that in the old quantum theory use was made of the classical conception of phase space. We can illustrate the meaning of this from the example of a particle moving along a direction x. Let us record in the way of classical physics pairs of values of x and of momentum p at x, and let us plot these points on a diagram as in figure 4. The area of the diagram, the points of which denote values of p and x, is the phase space of this particular case.

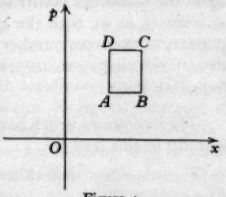


Figure 4

The quantum theory imposed the condition on the classical theory that the phase space could not be divided into rectangular areas of magnitude less than h.

In the light of Bohr's and Heisenberg's work we can now interpret this by saying that the physical world is so constituted that within a rectangle $ABCD$ which is equal to h it is impossible to distinguish one point from another. Thus we could have arrived at these uncertainty relations by means of a bold interpretation of Planck's theory without the intervention of waves. We should, however, have lost the complementarity of the particle and wave conceptions which the new quantum theory describes and explains.

But the fact that these relations can be so simply obtained serves to strengthen our belief in them as a physical principle. Once a wave description is employed the uncertainty relations follow from Rayleigh's theory of optical instruments. We might then ask whether the relations are merely brought in with the mode of description and whether they have any actual counterpart in the world of physics. Ultimately principles of physics rest upon experiment, and they are summaries of our observations. The uncertainty principle has not been verified by experiment, but an appeal has been made to an experiment of thought in the style of some well-known experiments with the engines of thermodynamics. There is, of course, nothing contrary to the methods of physics in such a procedure. The particular experiment we have in mind is that of the so-called "gamma-ray microscope" invented by Heisenberg. If we wish to locate a particle by an optical method as accurately as we can, we illuminate it with a very short wave-length and look at or photograph the object, using a microscope to help us. We know that the result of the attempt at location will be liable to be in error to the order of the wave-length of the light. In the case of an attempt to locate an electron we choose the shortest waves we can find for the purpose of illumination. This is why we speak of a gamma-ray microscope. (Perhaps we might use cosmic rays, but we are not sure whether they are waves or not.)

But there is a well-known effect known as the Compton effect which tells us that the act of illumination of the electron gives it a change of momentum of the order of magnitude h/λ which is undetermined in direction. The calculation shows that

this undetermined momentum Δp and Δx the indeterminacy in position of the electron are such that their product is of the order h. It will be noted that in this example an appeal has been made to two well-established physical effects and the result we have derived is thus in accordance with experiment. Other similar examples lead to the same result, so that it can be claimed that the principle is not simply introduced by the mode of description; the latter has indicated a result which appears to be in harmony with the results of observation.

It is natural that a principle of this character should have attracted attention by its originality and its importance. Much has been written upon it, and not all that has been written has been faithful to its physical character.

The principle must be regarded as a law about which all observers will agree, that is to say, that when observers have made experiments and eliminated their own vagaries and those of their apparatus, they will come to the conclusion that in particular circumstances they have been able to locate a particle to an accuracy Δx, and that it is then impossible to determine the momentum to a degree of accuracy greater than that prescribed by the relations (86).

The relation containing uncertainties in the time and energy measurements shows that if we attempt to measure the energy of a particle at an instant of time t, the conditions being such that a time interval Δt is required, then we cannot distinguish between energies W and $W+\Delta W$, where $\Delta W \Delta t \sim h$. This again gives the range which limits the application of the concept of particles with certain energies at certain times. A careful observer will reveal this uncertainty in measurements as part of the nature of the world of physics, he does not create it.

The indeterminacy of the pair of quantities is a consequence of the process of measurement which disturbs the phenomenon to be investigated. An example of this is the process of using a short-wave radiation to locate a particle. We cannot avoid the incompleteness in our measuring process which is described by the relations (86). Although we can change the magnitude of one member of the product, the product itself is always of the same order of magnitude.

In spite of the fact that the smallness of h leaves our macroscopic measurements with an apparent definiteness so that we can with great accuracy assign momentum to a particle at a point of space and energy at an instant of time, there is a fundamental change in principle implied by the discovery revealed in the case of minute measurements.

The scheme of classical physics assumed as the result of observation in the case of large-scale phenomena that absolutely accurate data were accessible. Thus with a complete knowledge of the data now it should be possible in a physical problem to trace the history of the process under examination and to anticipate its future course. The theory was deterministic and stress was laid on the possibility of prophecy which was thus inherent in it.

But the view is changed with the new quantum theory, and although it is natural to lay stress on the fact that it destroys the reputation of the prophet, that is only secondary to the critical difficulty, which is that we cannot know the present.

§ 8. THE WAVE FUNCTION

In the wave equations of physics the functions which satisfy them have some simple physical meaning like the displacement of a particle of a vibrating string, the pressure in a medium or the electric and magnetic intensities of a varying electromagnetic field. The meaning of the function ψ is different from that of the wave functions in these familiar cases.

We can give no more than a brief mention of the conclusions which have been arrived at in regard to its significance.

In the theory of vibrations of a string there are certain special sine and cosine functions which describe the states of the motion representing a fundamental mode or state and a series of overtones or higher states of vibration. A more complex state of vibration can be represented by means of a Fourier expansion by a summation of these simpler states. Each of the sine or cosine terms of the series may be said to represent a basic state.

In solving Schroedinger's equation we find that a series of values of ψ can be obtained with each of which is associated a particular value of W. We thus say that the basic states of the system are represented by a series of values ψ_1, ψ_2, etc., and that these are characterized by energy values W_1, W_2, etc.

Any more complex state is to be represented as a summation of these basic states, e.g.

$$\psi = \sum_n A_n \psi_n \qquad \text{.....(87).}$$

This form of representation of a system replaces the representation by models which was a feature of the classical and old quantum theories.

We rely upon the function ψ to give us probable or average values of quantities associated with the state described by it, and it is spoken of as a probability function. We also speak of the function ψ, since it satisfies an equation of wave form, as the probability wave.

We give only one example of the determination of a probable value. Suppose that a particle described by ψ lies in a certain region of space, then the probability that it lies within the element dv is given by $\psi\psi^* dv$, where ψ^* is the complex conjugate function corresponding to ψ.

Since the particle is certainly somewhere in the volume the integral $\int \psi\psi^* dv$ taken throughout the volume must be unity. This imposes a condition on the coefficients A_n and ψ is said to be normalized.

Difficulties associated with Schroedinger's equation. Difficulties were soon encountered in the original equation proposed by Schroedinger. It failed to give a satisfactory account of the fine structure of the hydrogen spectrum.

Sommerfeld's theory of the origin of fine structure was based on the use of the relativistic expression for mass instead of the constant value of Newtonian theory.

The natural procedure in wave mechanics is to modify the energy expression in Schroedinger's equation in the same way by introducing the relativistic form for the energy. But when this is done the expression for the energy levels, although similar in form to that of Sommerfeld, deviates from it in detail. Sommerfeld's

formula agrees very well with experiment in this case, so that the new values have to be rejected. Moreover, the difficulty pointed out by Millikan and Bowen to which we have referred still exists, and it is evident that some modification of the original equation is required.

A suggestion for a procedure which leads to a way out of this difficulty is found in an analogy suggested by Maxwell's equations. These equations form the basis of the electromagnetic theory of light. They are differential equations of the first order, and a wave equation of the second order satisfied by the components of the electric and magnetic intensities can be derived from them. This is the familiar wave equation of optics.

We may suppose that in the quantum theory the second-order equation has been discovered first, and that a deeper insight into the significance of the theory can be obtained from the discovery of the underlying first-order equations. It is in fact possible to attain the goal with the help of this analogy, but the whole idea appears to be of the nature of a mere analogy without any special significance. Moreover, Dirac obtained the equations originally by a quite different and more significant procedure.

If we consider the simplest case of a particle moving in the absence of a force the only energy is kinetic, which we can write in the form

$$W=\frac{1}{2m}(p_x^2+p_y^2+p_z^2) \qquad \text{.....(88).}$$

This is a case where the energy does not contain the time explicitly. Schroedinger's equation (62) for this case is

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}+\frac{8\pi^2 m}{h^2}W\psi=0 \qquad \text{.....(89).}$$

This may be written in the form

$$W\psi=\frac{1}{2m}\left\{\left(\frac{h}{2\pi i}\frac{\partial}{\partial x}\right)^2+\left(\frac{h}{2\pi i}\frac{\partial}{\partial y}\right)^2+\left(\frac{h}{2\pi i}\frac{\partial}{\partial z}\right)^2\right\}\psi \qquad \text{.....(90).}$$

The components of momentum of equation (88) thus become operators in equation (90), and a comparison of the two equations shows how the Hamiltonian equation of classical mechanics becomes the wave equation in the quantum theory. We can thus preserve the form of equation (88) and regard the components p_x, p_y and p_z as operators. Since W has, in this simple case, taken the place of $\partial S/\partial t$ of equation (34), it is evident that the energy W and the momentum p do not occur symmetrically either in the Hamiltonian expression or in the corresponding operator form taken into the quantum theory.

The wave equation (55) is symmetrical in the operators $\partial^2/\partial t^2$ and $\partial^2/\partial x^2$, etc., and equations (57) and (89) are special cases of equation (55), so that the wave equation is not faithful to its classical analogy. There is also a more serious objection for, if we put the Hamiltonian expression in a form which takes account of the dependence of mass upon velocity, the asymmetry between the time and space operators still remains. It becomes impossible to generalize the equation so that it

is satisfactory from the point of view of the theory of relativity. Thus Dirac was led to search for a new equation for a particle and not one which is simply a new form of the old one. The new equation is to contain W and the components p symmetrically, and it must be invariant with respect to the Lorentz transformations.

There is a very important energy relation of the special theory of relativity which forms the basis of this particular part of the discussion. If we denote the rest mass of a particle by m_0, its mass when moving with velocity v is $m_0/\sqrt{(1-v^2/c^2)}$. Since the energy is now kinetic, it follows that

$$\frac{W^2}{c^2} = m_0^2c^2 + p_x^2 + p_y^2 + p_z^2 \qquad \text{.....(91).}$$

If we convert the momentum components to operators as before and, in addition, convert W to

$$-\frac{h}{2\pi i}\frac{\partial}{\partial t},$$

the wave equation suggested by equation (91) expressed in operators is

$$(W^2/c^2 - m_0^2c^2 - p_x^2 - p_y^2 - p_z^2)\,\psi = 0 \qquad \text{.....(92).}$$

This is the wave equation in the form (55), and is the relativistic form of equation (89), becoming the latter when the velocity of motion is small compared with that of light.

But general requirements based upon the principle of superposition make it necessary that the fundamental equation of the quantum theory should be of the first order in the differential operator with respect to the time. Thus equation (92) must be a derived equation, and the equation taken by Dirac as the fundamental equation is

$$(W/c + \alpha_x p_x + \alpha_y p_y + \alpha_z p_z + \alpha_0 m_0 c)\,\psi = 0 \qquad \text{.....(93),}$$

where W, p_x, p_y and p_z are now the operators which we have introduced to replace these originally mechanical quantities.

The quantities α are matrices of a simple structure and they satisfy the relations

$$\alpha_m\alpha_n + \alpha_n\alpha_m = 0 \qquad \text{.....(94)}$$

and

$$\alpha_m^2 = 1 \qquad \text{.....(95).}$$

The suffixes m and n denote x, y, z or 0, so that equation (94) represents products of the matrices of equation (93), and in equation (94) m and n are unlike. If $m=n$ the relation (95) holds.

The equation which we have now obtained has the property that it is invariant with respect to the Lorentz transformation.

The result of this investigation is most striking when the equation is modified to include the case when the particle is charged and moves in an electromagnetic field. The procedure is a very simple one and follows from a recognition that the coordinate x and the momentum p_x are no longer conjugate quantities in this case. Instead of p_x the generalized momentum $\pi_x = p_x + e/c\,\phi_x$ is the conjugate of x, where

ϕ_x is a component of the electromagnetic potential and e is the charge associated with the particle. This means that the momentum p_x is replaced by

$$\pi_x - e/c\ \phi_x$$

in the equations, and the first-order equation is then obtained by replacing π_x by the operator $(h/2\pi i)\,(\partial/\partial x)$. There are similar modifications of p_y, p_z and W. In this way we obtain an equation which in form satisfies the requirements of the special theory of relativity. We might imagine that we have now progressed to the point of obtaining an equation which forms the basis of the theory of fine structure, taking the place of Sommerfeld's treatment of this subject. But it turns out that much more has been done, for Dirac has shown that the manipulation of the first-order equation which leads to the second-order equation contains certain additional terms over and above those occurring in Schroedinger's equation. The form in which they occur supplies the explanation of them. The second-order equation which is obtained by Dirac's procedure contains additional energy terms. These are absent from the Schroedinger equation, generalized to take account of the electromagnetic field and relativistic effects. The form of the terms suggests that the additional energy arises from a magnet of moment $he/4\pi m_0 c$, m_0 denoting the mass of the charged particle. This moment is that of the magneton associated with the theory of Bohr.

We must note how this magneton enters into Dirac's theory. It occurs merely from the recognition of the need to adopt at the basis of the quantum theory an equation satisfying in form the principle of restricted relativity. We are then led, without further assumption, to a result which may be described as associating a magnetic moment with the electron. We may remember how this association was made in the earlier theory. It was supposed that the electron had a degree of rotation about an axis, i.e. it had spin. This was the hypothesis of Goudsmit and Uhlenbeck.

The calculation of energy levels from Dirac's equation, which was carried out simultaneously by Darwin and Gordon, gives the correct fine-structure formula about which we spoke earlier in this report.

In order to be able to appreciate how some of the difficulties have been cleared up we will write down the expression for the energy levels given by the new method for the case of a charge in motion in a central field of force, i.e. with potential energy U equal to $-Ze^2/r$:

$$W = m_0 c^2 \left[1 + \frac{\alpha^2 Z^2}{\{n_r + \sqrt{(k^2 - \alpha^2 Z^2)}\}^2}\right]^{-\frac{1}{2}} \qquad \text{.....(96).}$$

In this formula Z denotes the number of charges situated on the nucleus about which the electron of energy W is moving. n_r is an integer which in the older form of the quantum theory was called the radial quantum number and k the azimuthal quantum number. α is the fine structure constant $2\pi e^2/hc$.

Although the significance of the quantum numbers in the new theory is not the same as in the older, it is convenient to describe them in the same way.

We shall describe the number

$$n = n_r + k \qquad \text{.....(97)}$$

as the total quantum number, but k can now have positive or negative values and equation (97) must be replaced by

$$n = n_r + |k| \qquad \ldots\ldots(98).$$

Thus n depends only upon the numerical value of k, not upon its sign.

Different values of k with the same value of n give the relativity doublets. Thus equation (96) is identical, to this extent, with Sommerfeld's formula. But the double character of the formula due to the fact that k can have negative as well as positive values remains concealed. It comes into expression when the central coulomb field is modified, as when the inner electrons exert a shielding effect on an outer electron, and it becomes evident in the case where a magnetic field is present.

These finer considerations require no additional hypotheses such as were required to be included in the older theory. The atomic model with its orbits and electron spins is replaced by Dirac's equation which places the account of those phenomena associated with the region outside the nucleus of the atom on a satisfactory quantitative basis.

References

(13) Heisenberg. *Z. Phys.* **33**, 879 (1925).
(14) Born and Jordan. *Z. Phys.* **34**, 858 (1925).
(15) Born, Heisenberg and Jordan. *Z. Phys.* **35**, 557 (1926).
(16) Dirac. *Proc. roy. Soc.* **109**, 642 (1925); **110**, 561 (1926).
(17) Dirac. *Quantum Mechanics*, 2nd ed. Oxford: Clarendon Press.
(18) Frenkel. *Wave Mechanics, Advanced General Theory*. Oxford: Clarendon Press.
(19) Sommerfeld. *Atombau u. Spektrallinien Wellenmechanisher*, Ergänzungsband. Vieweg und Sohn.

ATOMIC PHYSICS

By C. H. COLLIE, M.A., B.Sc., Clarendon Laboratory, Oxford.

§ 1. INTRODUCTION

The year has been one of steady progress in which no spectacular advances have been made. Although the large output of short papers announcing the discovery of new radio elements continues, there is a welcome tendency to study the changes already known in greater detail and with the highest possible precision. So far this procedure has always yielded results of great interest and importance. The application of the new mechanics to nuclear problems has so far met with uninterrupted success in the sense that no phenomena are known of which it cannot give a qualitative explanation. As, however, both the nature of the forces holding nuclear particles together and the numerical constants describing them are at our disposal and can be varied within wide limits, it is difficult to assess the value of these initial successes. There has always been some doubt whether quantum mechanics can be applied to particles of very high energy, and Blackett and Wilson's measurements of the energy lost by electrons in a lead plate show that, for energies greater than 10^9 ev., electrons are much more penetrating than was expected. The theory of positron emission also offers some difficulties; the absorption of extra-nuclear electrons by nuclei capable of emitting a positron appears to be less frequent than theory predicts, and there seems little doubt that in the next few years it will be possible to subject our present theoretical ideas to exacting experimental tests.

A brief perusal of the *Report* will show the remarkable extent to which investigators are devoting themselves to the study of the nucleus. Even the spectroscopist is chiefly interested in hyperfine structure, an essentially nuclear phenomenon, while the technique of low-temperature measurement has been pressed into service to measure the magnetic moment of the proton.

In conclusion it may not be out of place to draw attention to an exhaustive report on nuclear physics issued in three volumes by the American Physical Society[1], in which a full discussion of the theoretical aspects of the subject will be found.

Reference

(1) *Rev. Mod. Phys.* **8** *and* **9**.

§ 2. ISOTOPES

(*a*) *New mass data.* As reported last year, mass spectroscopic measurement is regaining that degree of precision and internal consistency generally attributed to it before accurate measurements of nuclear disintegration energies were made. The only outstanding discrepancy of any importance is the small difference between the American and English workers' values for the mass of ^{13}C.

Aston[2] has redetermined the isotopic weight of carbon with very concordant results by measuring no fewer than eight different doublets, namely:

$$O—CH_4,\ CO—C_2H_4,\ O^{++}—C\sim A^{+++}—A^{++},\ ^{36}A—C_3,\ O^{++}—C\sim C—OH_2,$$
$$He—C^{++}\sim A^{+++}—A^{++} \text{ and } ^{40}A—C_2H_4.$$

He has obtained $12{\cdot}00355 \pm 0{\cdot}00015$ for the isotopic weight of carbon, a value in complete agreement with the value 12·0035 obtained from disintegration data. A new determination has also been made by Bainbridge and Jordan[3] who obtain $12{\cdot}00428 \pm 0{\cdot}00017$ from one set of four doublets and $12{\cdot}00402 \pm 0{\cdot}00017$ from measurements of doublets formed by the hydrides of the light elements.

The mean value 12·00415 exceeds Aston's value by 0·00060, i.e. by an amount which is several times the estimated experimental error.

No explanation of this difference has yet been given; as table 1 shows the agreement between the values obtained for other elements is all that could be desired.

Table 1

Atom ...	^{1}H	^{2}D	^{4}He	^{10}B	^{12}C	^{20}Ne	^{40}A
Aston	$1{\cdot}0081_2$	$2{\cdot}0147_1$	4·0039	$10{\cdot}0161$	$12{\cdot}003_5$	$19{\cdot}998_4$	$39{\cdot}975_4$
Bainbridge and Jordan	$1{\cdot}0081_5$	$2{\cdot}0147_8$	$4{\cdot}0039_8$	$10{\cdot}016_2$	$12{\cdot}004_0$	19·994	$39{\cdot}975_8$

Pollard[4] and his coworker have been able to estimate the masses of some rare isotopes by measuring the maximum range of the protons emitted when the lighter elements are bombarded by α particles. The quantity actually measured is, of course, the energy release (E) in a change of the type

$$^{35}_{17}Cl + {}^{4}_{2}He \rightarrow {}^{38}_{19}A + {}^{1}_{1}H + E,$$

so that isotopic masses obtained in this way are not absolute but also depend upon measurements with a mass spectrograph. With the exception of ^{25}Mg, ^{37}S and ^{39}K exact masses are known now for all the isotopes below ^{40}A.

If these exact isotopic masses[5] are plotted against the nearest whole number the series of minima observed by Oliphant[6] for the light elements is found to extend in a less pronounced form up to ^{40}A.

A modification of this method of plotting mass data has been used by Bothe and Maier-Leibnitz[7]. They select only those elements which have a given number of excess neutrons, but include both stable and unstable elements on the same curve.

In this way a series of smooth curves is obtained from which the masses of unknown isotopes can be obtained by interpolation. They are able to predict that both ^{5}Li and ^{5}He will have such large masses that they will break up at once, but that $^{4}_{1}H$ will be a very short-lived β-radioactive substance. As will be seen later in the *Report* there is a little evidence that ^{5}He does in fact break up as soon as it is formed; $^{4}_{1}H$ is of course still unknown.

(*b*) *Nuclear moments.* No new method of measuring nuclear properties has been developed during the year, but the methods already available continue to be applied to new nuclei.

The magnetic moment of the proton has been determined by a direct macroscopic measurement of the susceptibility of hydrogen at low temperatures. The diamagnetic susceptibility of normal hydrogen decreases owing to the increasing paramagnetism of the nucleus down to 1·78° K., the lowest temperature at which it has been measured. The rotational moment of the ortho-molecule remains down to the lowest temperature. By taking Frisch and Estermann's[9] value of 0·85 for its molecular moment a value of $\mu_p = 2{\cdot}3 \pm 0{\cdot}2$ is obtained in fair agreement with the results obtained by other methods.

The new method of measuring the sign of nuclear magnetic moments described in last year's *Report* has been applied to ^{7}Li, ^{23}Na, ^{39}K[11], ^{85}Rb and ^{87}Rb[10]. All these are found to have a positive moment in agreement with the spectroscopic data. The agreement in the case of potassium is particularly gratifying, as it was originally reported to have a negative moment[12]. Potassium has a very small nuclear magnetic moment, and its hyperfine structure was only observed with great difficulty. While the scale of the hyperfine structure is determined by the magnitude of the magnetic moment, it is its sign which determines the relative intensity of the components. In the original measurements the highest available resolving power was used; this yielded plates which while well adapted to measurement of the hyperfine structure were ill suited to intensity measurements owing to the use of a long path difference which allowed light from the background (the lines were observed in absorption) to spread into the absorption lines.

New measurements[13] using a smaller path difference show that ^{39}K has a positive magnetic moment.

Magnetic moments have been measured by the spectroscopic method for Sc $(+4{\cdot}8)$[14], ^{67}Zn $(+0{\cdot}9)$[15], ^{115}In $(5{\cdot}3)$[17] and Tl $(1{\cdot}45)$[18]: also mechanical moments for ^{27}Al $(9/2)$[16] and Cs $(7/2)$[19]. Departures from the interval rule continue to be interpreted as nuclear quadrupole moments, and values for this nuclear constant have been obtained for ^{185}Re[20] $(q = 2{\cdot}6 \times 10^{-24})$, ^{157}Re $(q = +2{\cdot}6 \times 10^{-24})$ and ^{115}In $(q = +0{\cdot}8 \times 10^{-24})$[21].

In conclusion, attention may be called to a general article by Sexl[22] dealing simply and clearly with the theoretical principles underlying the determination of nuclear constants.

(*c*) *Transuranic elements*. The investigation of the elements with a larger atomic number than uranium which has been carried out during the past three years at the Kaiser Wilhelm Institut is now sufficiently complete for a brief account of it to be given here[23]. Contrary to the view originally expressed[24], no α particles are emitted when uranium is bombarded with neutrons, the neutron being captured to form another uranium atom of short period. The three following radioactive series have been established:

(1) $U + n \rightarrow {}^{10''}_{92}U \xrightarrow{\beta} {}^{2{\cdot}2'}_{93}\text{eka Re} \xrightarrow{\beta} {}^{56'}_{94}\text{eka Os} \xrightarrow{\beta} {}^{66\text{ hr}}_{95}\text{eka Ir} \xrightarrow{\beta} {}^{2{\cdot}5\text{hr}}_{96}\text{eka Pt} \xrightarrow{\beta} {}_{97}\text{eka Au}.$

(2) $U + n \rightarrow {}^{40''}_{92}U \xrightarrow{\beta} {}^{16'}_{93}\text{eka Re} \xrightarrow{\beta} {}^{5{\cdot}7\text{ hr}}_{94}\text{eka Os} \xrightarrow{\beta} {}_{95}\text{eka Ir}.$

(3) $U + n \rightarrow {}^{27'}_{92}U \xrightarrow{\beta} {}_{93}\text{eka Re}.$

The first two reactions go with either fast or slow neutrons, but the third reaction is a typical resonance capture reaction which takes place with slow neutrons only. Comparison with boron shows a sharp resonance at 25 ± 10 ev. The absorption coefficient (μ/ρ) in uranium nitrate is 3 cm^2/g.

Consideration of the capture cross-section leads to the conclusion that the absorbing atom must be the principal isotope $^{238}_{92}U$. If $^{235}_{92}U$ were concerned the cross-section at resonance would be 3×10^{-18} cm^2, a value which is larger than $\lambda^2/4\pi$, the maximum value allowed by theory. Similarly the capture cross-sections observed for fast neutrons in the first two changes are approximately the same and equal to $1{\cdot}6 \times 10^{-25}$ cm^2, a value which again indicates that the principal isotope $^{238}_{92}U$ is the capturing atom. Thus one is forced to the conclusion that all three β radioactive uranium atoms have the same mass and are therefore isomers.

The phenomenon of nuclear isomerism is not new, but has been neglected owing to the reluctance of theoretical physicists to admit that the state of a nucleus is not uniquely determined by its mass and charge. Apart from the classical example of uranium Z and uranium X_2 several isomeric pairs have been suspected amongst the lighter radio elements.* The recent investigations of Bothe and Gentner[25] on photo-nuclear disintegration have greatly increased the probability that isomerism is the explanation of the observed anomalies. The case of bromine is particularly clear. Irradiation of bromine by slow neutrons produces three types of radioactive bromine atoms of periods P_1, P_2, P_3. Irradiation by $17{\cdot}10^{+6}$ ev. γ rays also produces three types of radioactive bromine with periods P_2, P_3, P_4 by the ejection of a photoneutron. A careful investigation shows that bromine has only two isotopes, ^{79}Br and ^{81}Br. Since P_2 and P_3 are formed both by the addition to, and by the removal of a neutron from bromine, they must both be ^{80}Br, so that P_1 may be identified with ^{82}Br and P_4 with ^{79}Br. Other examples of the phenomenon are afforded by indium and gallium. It is interesting to note that as far as present observations go[27] the β-ray spectra of ^{80}Br are identical, so that the difference in period must be entirely accounted for by differences in the nuclear configuration of the two nuclei.

(*d*) *Separation of isotopes.* A great deal of experimental work on the separation of isotopes has been carried out recently, and nearly all the available methods† except the thermal diffusion method of Chapman[28] have been used successfully. The classical method of separating pure radioactive isotopes has been applied to rubidium[29]. A kilogram of a Canadian mica containing a few per cent of rubidium was worked up chemically and yielded 250 mg. of Sr which, when examined in a mass spectrograph[30], was found to be pure ^{87}Sr (ordinary strontium is, of course, nearly all ^{88}Sr) formed throughout geological time by the β disintegration of ^{87}Rb[30a]. The isotope responsible for the radioactivity of potassium has been separated in a mass spectrograph and identified as ^{40}K by Smythe and Hemmen-

* The example reported by Libby (see last year's *Report*) has been reinvestigated by Roaf[26] who has been unable to confirm it.

† A new method has been suggested by Stern but it has not yet been tried. It consists in balancing the gravitational deflection of an atomic beam by the magnetic deflection due to its magnetic moment in an inhomogeneous field.

dinger[31]. Thus the at one time widely held view that the radioactivity of the alkali metals was due to small quantities of a radioactive homologue $_{87}$eka Cs is finally disposed of. A further search for this element has been made by Hirst[32] who examined the X radiation emitted by some Cs residues from lepidolite. A faint line in the expected position for Kα was observed. The line was, however, diffuse and disappeared when the crystal was rocked over a somewhat larger arc. By a fortunate chance the calcite crystal was the one used by Papish and Wainer[33] in a similar search some years ago; the unconfirmed positive result which they obtained may well have been due to the use of this imperfect crystal.

The diffusion method as developed by Herz[34] has been successfully[35] employed to raise the proportion of ^{36}A in argon from 0·3 to 50 per cent and to obtain a considerable concentration of the heavy isotopes of carbon and nitrogen[36]. The most generally applicable method of separating isotopes is centrifuging[37]. Its great advantage is that the degree of separation depends upon Δm, the difference in mass of the isotopes, and not as in other methods upon $\Delta m/m$, the relative mass difference. It is therefore as effective a method for separating the isotopes of mercury as for separating those of hydrogen. For this reason it is easier to choose a suitable compound with which to carry out the separation.

Hitherto it has not been possible to construct a centrifuge capable of withstanding the disruptive forces associated with the necessary peripheral speed of 10^5 cm./sec. The development of light alloys for aeroplane construction has encouraged several workers to attempt this method of separation, and Beams and Masket[38] claim to have separated the isotopes of chlorine by a combined process of evaporation and centrifuging. An enclosed metal rotor containing carbon tetrachloride was evacuated through a hollow central axle: carbon tetrachloride vapour was pumped off and condensed while the centrifuge was running and the density of the separate fractions measured. As the separation has not yet been checked in a mass spectrograph and an increase in density could easily be produced by a very small amount of impurity of high molecular weight, the result must be accepted with caution.

Another method capable of wide application is the photo-chemical one which has been applied to mercury by Zuber[39]. When illuminated by its resonance radiation 2537 A. ($2\ ^3P_1$–$1\ ^1S_0$) mercury vapour is activated and combines with oxygen to form an oxide which is deposited. Since the isotopes have different absorption spectra (due to differing hyperfine structure) it should be possible to activate each isotope in turn with the appropriate radiation and achieve a complete separation by collecting the mercury oxide formed. A method of obtaining the necessary monochromatic light had already been suggested by Mrozowsky[40], who showed that the transverse Zeeman pattern of mercury can be adjusted (by a suitable choice of field) to coincide with a large number of the components of the resonance line.

Using such a filter Zuber (who followed the reaction optically) found that a complete separation was not attained owing to complications in the subsequent chemical reaction. The efficiency of the separation was increased by the addition

of nitrogen, and there is little doubt that the method could be developed into a practical method of separating isotopes. It is interesting to note that the quantities of matter involved (5×10^{-8} g.) are of the same order as were used by Oliphant in his bombardment of the separated lithium isotopes.

Electro-chemical methods of separating the lithium isotopes have been developed by Lewis[41], who dropped a fine spray of lithium amalgam down a column of lithium chloride dissolved in ethyl alcohol (this solvent was used to avoid the formation of hydrogen). After descending the column, which was 18 m. high, the amalgam was collected and converted into lithium chloride, which was added to the ascending electrolyte. After 10 l. of amalgam had passed through the apparatus the amalgam at the foot of the column was removed and the atomic weight of the lithium in it determined. The ratio of Li_7/Li_6 was found to have decreased from 11·6 to 5·1. The separation was confirmed with a mass spectrograph and corresponds to a difference of electrode potential of 6×10^{-4} v. between the isotopes.

References

(2) Aston. *Nature, Lond.*, **139**, 922 (1937).
(3) Bainbridge and Jordan. *Phys. Rev.* **51**, 384 (1937).
(4) Pollard and Brasefield. *Phys. Rev.* **51**, 8 (1937).
(5) Pollard and Brasefield. *Nature, Lond.*, **137**, 943 (1936).
(6) Oliphant. *Nature, Lond.*, **137**, 396 (1936).
(7) Bothe and Maier-Leibnitz. *Naturwissenschaften*, **25**, 26 (1937).
(8) Laserew and Schubnikow. *Phys. Z. Sowjet.* **10**, 117 (1936).
(9) Frisch and Estermann. *Z. Phys.* **85**, 4, 17 (1933); **86**, 132 (1933).
(10) Milman and Zacharias. *Phys. Rev.* **51**, 1049 (1937).
(11) Kelogg, Rubi and Zacharias. *Phys. Rev.* **50**, 472 (1936).
Torrey. *Phys. Rev.* **51**, 501 (1937).
(12) Jackson and Kuhn. *Nature, Lond.*, **137**, 108, (1936).
(13) Jackson and Kuhn. *Nature, Lond.*, **140**, 276 (1937).
See also Meissner and Luft. *Z. Phys.* **107**, 762 (1937).
(14) Kopfermann and Wihke. *Z. Phys.* **105**, 16 (1937).
(15) Lyshede and Rasmussen. *Z. Phys.* **104**, 468 (1937).
(16) Jackson and Kuhn. *Nature, Lond.*, **140**, 110 (1937.)
(17) Schüler and Schmidt. *Z. Phys.* **104**, 468 (1937).
(18) Schüler and Schmidt. *Z. Phys.* **104**, 468 (1937.)
(19) Fölsche. *Z. Phys.* **105**, 133 (1937).
(20) Schüler and Schmidt. *Z. Phys.* **104**, 468 (1937).
See also Bacher and Tomboullin. *Phys. Rev.* **50**, 1096 (1936).
(21) Schüler and Korsching. *Z. Phys.* **105**, 168 (1937).
(22) Sexl. *Naturwissenschaften*, **25**, 153 (1937).
(23) Meitner, Hahn and Strassmann. *Z. Phys.* **106**, 249 (1937).
(24) Meitner, Hahn and Strassmann. *Ber. dtsch. chem. Ges.* **69**, 905 (1936).
(25) Bothe and Gentner. *Z. Phys.* **106**, 236 (1937).
(26) Roaf. *Proc. roy. Soc.* A, **159**, 133 (1937).
(27) Alichanian, Alichanow, and Ożelepow. *Phys. Z. Sowjet.* **10**, 78 (1936.)
(28) Chapman. *Phil. Mag.* **38**, 182 (1919).
(29) Hahn, Strassmann and Walling. *Naturwissenschaften*, **25**, 189 (1937).
(30) Mattauch. *Naturwissenschaften*, **25**, 156 (1937).
(30*a*) Smythe and Hemmendinger. *Phys. Rev.* **51**, 1052 (1937).
(31) Smythe and Hemmendinger. *Phys. Rev.* **51**, 178 (1937).
(32) Hirst. *Phys. Rev.* **51**, 584 (1937).

(33) PAPISH and WAINER. *J. Amer. chem. Soc.* **53**, 3818 (1931).
(34) HERZ. *Z. Phys.* **79**, 108 (1932).
(35) BARWICK and SCHÜTZE. *Z. Phys.* **105**, 395 (1937).
(36) WOOLRIDGE and SMYTHE. *Phys. Rev.* **50**, 233 (1936).
(37) ASTON. See *Mass Spectra and Isotopes* for a full discussion.
(38) BEAMS and MASKET. *Phys. Rev.* **51**, 384 (1937).
(39) ZUBER. *Helv. phys. Acta*, **9**, 285 (1936).
(40) MROZOWSKY. *Z. Phys.* **78**, 826 (1932).
(41) LEWIS. *J. Amer. chem. Soc.* **58**, 2519 (1936.)

§ 3. NUCLEAR TRANSFORMATIONS

It is not proposed to report at length on the large volume of work which has been carried out in this rather specialized field. This aspect of atomic physics has been fully reported during the last two years, and no fundamental advances have been made during the period under review. A large amount of important qualitative work has been carried out using the most effective particles available (6×10^6 ev. deuterons) which can produce nuclear changes in most elements irrespective of their atomic number. However, most of the accurate quantitative measurements continue to be made upon the disintegration of light elements using incident particles of moderate energy.

A large number of elements have been bombarded with fast deuterons and the nuclear changes taking place roughly identified. An interesting example is nickel[42], which when activated by 5×10^6 ev. deuterons forms a radioactive copper which decays to half value in 3·4 hr. and emits positrons.

The change involved is probably

$$^{60}_{28}\mathrm{Ni} + {}^{2}_{1}\mathrm{D} \rightarrow {}^{61}_{29}\mathrm{Cu} + {}^{1}_{0}n,$$

but the alternative process of deuteron capture by $^{58}_{28}$Ni cannot be excluded. As is to be expected, the excitation function of this reaction in which the charge of the nucleus is changed has a different shape from that of the typical rejection reaction which occurs when copper is similarly bombarded.

Further study of the reaction between lithium and deuterium has yielded results of great interest. Experiments with pure ^{7}Li[43] targets show that the protons emitted in the reaction

$$^{7}\mathrm{Li} + {}^{2}\mathrm{D} = {}^{8}\mathrm{Li} + {}^{1}\mathrm{H}$$

have a maximum range of 8 cm. and not 26 cm. as previously supposed. The low energy of the emitted proton leaves a considerable proportion of the available energy (calculated from mass data) unaccounted for: further examination[45] of the radio lithium (^{8}Li) produced shows that it emits α particles as well as the previously observed β particles. The following facts have been established by several workers: (*a*) The half-life period ($\sim$0·9 sec.) is the same whichever disintegration particle is observed. (*b*) The excitation curve for the production of this induced radioactivity is the same for both α and β particles and is different from other lithium excitation functions. (*c*) Approximately equal numbers of α and β particles are produced.

These experimental results can be explained by assuming the intermediate formation of an excited $^8Be^*$ nucleus which can subsequently disintegrate into two α particles:

$$^8Li \rightarrow {}^8Be^* + e^- : {}^8_4Be^* \rightarrow {}^4_2He + {}^4_2He.$$

The α particles do not form a homogeneous group but have a wide energy distribution extending from 0·6 to 6×10^6 ev. with a sharp maximum at $1{\cdot}3 \times 10^6$ ev. In order to account for this wide range of distribution the excited $^8Be^*$ nucleus would have to be formed either in several excited states or in a state of very ill-defined energy. In view of this difficulty and their failure to observe these α particles in other reactions which might be expected to give rise to $^8Be^*$ (i.e. the bombardment of ^7Li with protons or ^6Li with deuterons) Rumbaugh, Roberts and Hafsted[45c] suggest that both the α particle and the β particle are emitted in a complicated four-body disintegration:

$$^8_3Li \rightarrow {}^4_2He + {}^4_2He + e^- + {}^0_0n.$$

The proposed change has to be made a four-body reaction by including a neutrino (^0_0n) in order to reconcile the very small number of high-energy α particles observed with the characteristic shape of the β-ray spectrum of ^8Li. It is not possible to decide between these conflicting views until more data are available, but there is considerable evidence† that very short-lived intermediate products do play a considerable part in determining the distribution of energy between the nuclear fragments in reactions such as this in which the nuclei are broken up into α particles.

The same reaction (^7Li, D) has also been investigated to see if any ^5_2He was formed by the following nuclear change:

$$^7_3Li + {}^2_1D \rightarrow {}^4_2He + {}^5_2He. \qquad \text{.....}(a).$$

This change should manifest itself by the occurrence of two homogeneous groups of short-range α particles superimposed upon the general background which consists of particles with a continuous energy spectrum. The first attempts[46] to detect the group gave negative results, showing that less than one-fiftieth of the nuclear disintegrations produced proceed according to equation (*a*). Later workers[47] claim to have found a single group of α particles with a range of 7·1 cm.: their origin is obscure, since reaction (*a*) should give rise to two groups of α particles of equal momentum but different ranges. If the single group observed is identified with ^5_2He, then energy considerations show that ^5He is unstable by $9{\cdot}3 \times 10^5$ ev. If the group is due to ^5He there should be another group of range 11·7 cm. which was looked for but not found. A possible explanation of these results is that ^5_2He has a sufficiently long life to give the ^4He particles a definite range, but subsequently breaks up into ^4_2He and a neutron.

References

(42) Thornton. *Phys. Rev.* **51**, 893 (1937).
(43) Rumbaugh and Hafsted. *Phys. Rev.* **50**, 681 (1936).
(44) Gilbert, Smith and Fremlin. *Nature, Lond.*, **138**, 841 (1936).

† Cf. the succeeding paragraph; also last year's *Report*, p. 79.

(45) (*a*) LEWIS, BURCHAM and CHANG. *Nature, Lond.*, **139**, 24 (1937).
(*b*) RIDENOUR and HENDERSON. *Phys. Rev.* **51**, 1102 (1937).
(*c*) RUMBAUGH, ROBERTS and HAFSTED. *Phys. Rev.* **51**, 1106 (1937).
(46) KEMPTON, BROWNE and MANSDORF. *Proc. roy. Soc.* A, **157**, 372 (1936).
(47) WILLIAMS, SHEPHERD and HOXBY. *Phys. Rev.* **51**, 888 (1937).

§ 4. THE PASSAGE OF RADIATION THROUGH MATTER

(*a*) *Gamma rays.* During the last year our knowledge of the photo-nuclear effect has been greatly extended. This effect, which consists of the capture of a γ ray and subsequent emission of a neutron by a nucleus, was hitherto known to occur only with beryllium and heavy hydrogen, because they are the only elements which have a binding energy (for neutron emission) less than that of the readily accessible γ rays from Ra C′ and Th C′. By using the $1{\cdot}7 \times 10^7$ ev. γ rays which are emitted when lithium is bombarded with protons Bothe and Gentner[25] have been able greatly to extend the range of elements in which the photo-nuclear effect can be observed.

The neutrons liberated by the action of γ rays upon beryllium and heavy hydrogen are emitted in sufficient quantity for the neutron to be detected directly by means of the radioactivity it produces in other elements when it is captured. The number of neutrons produced from the heavy elements is much too small to be detected directly (because of the relative weakness of the γ ray source), and the photo-nuclear effect can only be observed for those elements in which the residual nucleus is radioactive.

Apart from its intrinsic importance the production of radio elements by neutron loss has greatly increased the certainty with which masses can be assigned to the artificially produced radioactive nuclei. A single example will be sufficient to indicate the method employed.

Silver has two stable isotopes, ^{107}Ag and ^{109}Ag, which can each capture a slow neutron and form a radioactive silver atom. These are distinguished by their periods (23 sec. and 2·3 min.), but it has hitherto been impossible to tell which is ^{108}Ag and which ^{110}Ag. When exposed to γ rays silver again forms two radioactive silvers with characteristic half-life periods (24 min. and 2·3 min.), but this time by neutron loss. It is evident that Ag (2·3 min.), which is formed both by neutron loss and by neutron capture, must be identified with ^{108}Ag formed by the reactions

$$^{107}\mathrm{Ag} + {}^{1}_{0}n \rightarrow {}^{108}\mathrm{Ag}; \quad {}^{109}\mathrm{Ag} + h\nu \rightarrow {}^{108}\mathrm{Ag} + {}^{1}_{0}n.$$

Thus mass numbers can definitely be assigned to all the five silver isotopes known, namely, ^{106}Ag (24 min.), ^{107}Ag (stable), ^{108}Ag (2·3 min.), ^{109}Ag (stable), ^{110}Ag (23 sec.). As can be seen in table 2, there is satisfactory agreement between masses assigned in this way and those obtained by applying similar reasoning to radio elements produced by fast neutron bombardment, a process which can also give rise to neutron emission.

So far the only photo-nuclear reaction which has been observed is the ejection of a single neutron. Bothe and Gentner looked carefully for other reactions but could not find them.

For example, Cr^{53}, Fe^{57}, Pd^{111}, Sn^{117}, Ba^{125}, Pb^{207} would all form known, easily recognizable radio elements if they expelled a proton. Thus, cadmium would form the silver isotope ^{110}Ag (23 sec.) already discussed:

$$Cd^{111} + h\nu \rightarrow Ag^{110} + {}^{1}H.$$

As mentioned above, there were no indications that such (γ, H) processes took place.

The observed cross-sections are in reasonable agreement with theoretical predictions; by far the largest cross-section observed was that of copper ($\sigma = 5 \times 10^{-28} cm^2$), a value which may be compared with that found for Be and Ra C′ γ rays (5×10^{-28} cm^2)[48].

As is to be expected for a photo effect, the cross-section varies capriciously from element to element, and many changes which might have been expected to occur were not observed to take place under the conditions of the experiment. Thus it was not found possible to produce detectable quantities of the known elements ^{13}N, ^{15}F, ^{28}Al, ^{27}Si, Cl^{34}, Cu^{64}, Zn^{65}, Pd^{107} by irradiating the stable element with the γ rays from lithium bombarded with protons.

A brief discussion of the theoretical points involved will be found in the ninth volume of the *Review of Modern Physics*[1].

(*b*) *Neutrons.* It is only intended to discuss in this section the passage of fast neutrons through matter. The numerous and complicated phenomena which have been observed when slow neutrons are used are treated in a separate section by Dr Moon.

The early work[49] using fast neutrons showed that as well as producing nuclear transformations they suffered inelastic collisions. The detailed investigation of the changes concerned has proved unexpectedly difficult, and we are still without any clear understanding of the processes involved. Using the $2{\cdot}4.10^6$-ev. neutrons from a D.D source, Kikuchi*[50] and his coworkers have shown that nearly all elements emit γ rays whose energy is approximately the same as that of the incident neutrons. The cross-sections for γ ray production when plotted against atomic number are found to lie roughly on a smooth curve which has a flat maximum at $z = 60$ and then drops sharply to a value comparable to the cross-sections found for the light elements[51].

The transmission of neutrons from a (Ra-Be) source through lead, iron and copper has been investigated by Grahame and others[52]. When the neutrons were detected by the γ rays which they produced in lead[53] absorption, cross-sections in the neighbourhood of 10^{-24} cm^2 were observed. The energy of the emitted γ quanta was estimated to be 10^6 ev., but it was not possible to find out the energy of the neutrons which were responsible for the emission.

The collision cross-section of neutrons and protons has been measured by several workers using neutrons of different energy[54]. The results obtained are all in excellent agreement with the current theory of the deuteron[55]. The angular

* The suggested transformation of the neutron into an electron and a proton has not been confirmed[44].

distribution of the scattered neutrons[56] has been reinvestigated theoretically using such a deuteron model with a nuclear radius equal to $2{\cdot}7 \times 10^{-13}$ cm. and a triplet depth $36{\cdot}04 \times 10^6$ ev. corresponding to a $2{\cdot}2 \times 10^6$ ev. binding energy for the deuteron.

The neutron energy at which the departure from symmetrical scattering (in coordinates in which the centre of gravity is stationary) becomes noticeable depends chiefly upon the degree to which the potential hole is rounded off. No accurate experimental data are yet available to compare with the theory, but some preliminary experiments by Harkins and others[57] using a RdTh-Be source indicate that more protons are scattered forward than corresponds with symmetrical scattering.

One of the disintegration processes discussed by Fermi in his first paper on induced radioactivity was the ejection of a neutron by an $(n, 2n)$ reaction from the activated nucleus. By using a powerful artificial source of fast neutrons (Li + D) it has been shown that such reactions do in fact take place: indeed, one example has been reported in which two neutrons are ejected simultaneously[58]. When scandium is bombarded by the fast neutrons produced when 6×10^6 ev. deuterons fall on lithium, two positron emitting scandium isotopes are produced. The same isotopes are produced by bombarding calcium with deuterons[61] and potassium with α particles. The interpretation of this phenomenon given by the authors is reproduced below:

$$^{45}_{21}\mathrm{Sc} + {}^{1}_{0}n \rightarrow {}^{44}_{21}\mathrm{Sc} + 2\,{}^{1}_{0}n;\ {}^{44}_{21}\mathrm{Sc} \rightarrow {}^{44}_{20}\mathrm{Ca} + e^{+} \quad (52 \text{ hr.}).$$

$$^{45}_{21}\mathrm{Sc} + {}^{1}_{0}n \rightarrow {}^{43}_{21}\mathrm{Sc} + 3\,{}^{1}_{0}n;\ {}^{43}_{21}\mathrm{Sc} \rightarrow {}^{43}_{20}\mathrm{Ca} + e^{+} \quad (4 \text{ hr.}).$$

A wide variety of these reactions has been investigated by Heyn[59] and others[60], and the isotopes produced in this way and by photo-disintegration are collected in table 2. All the radiations have not been examined in detail, but those elements marked with an asterisk have been shown to be positron emitters.

Table 2. Induced radioactivity due to neutron-loss

(*a*)	$_{7}$N	$_{8}$O	$_{15}$P	$_{29}$Cu	$_{30}$Zn	$_{31}$Ga	$_{34}$Se
(*b*)	^{13}N	^{15}O	—	^{62}Cu, ^{64}Cu	—	^{68}Ga	Se (56 min.)
(*c*)	—	^{15}O	^{30}P	^{62}Cu	^{63}Zn	^{67}Ga, Ga^{70}	Se (56 min.)

(*a*)	$_{35}$Br	$_{47}$Ag	$_{49}$In	$_{51}$Sb	$_{52}$Te	$_{80}$Hg	$_{81}$Tl
(*b*)	Br^{78}	^{106}Ag*, ^{108}Ag	^{112}In	^{120}Sb* (13 min.)	Te (60 min.)	Hg (43 min.)	^{201}Tl
(*c*)	^{78}Br, ^{80}Br (16 min.)	^{106}Ag, ^{108}Ag	^{112}In	^{120}Sb	Te (60 min.)	—	—

(*a*) Element irradiated
(*b*) Radioactive nucleus produced by fast neutrons
(*c*) Radioactive nucleus produced by γ rays

No systematic study of the phenomenon has yet been made, but in general the cross-section for the $(n, 2n)$ process is larger than the photo-nuclear cross-section.

So far the ejected neutron has not been detected, and it is not known how the available energy and momentum is divided between the two neutrons.

(*c*) *Electrons.* Experiments on the passage of fast electrons through matter are of great importance at the present time. Theory shows that fast electrons when passing through an atom lose energy chiefly by radiation and not by ionization. As was pointed out by Bethe and Heitler[63] there is reason to believe that electrons with energy in the neighbourhood of 10^9 ev. lose very much less energy than quantum theory predicts, and that this is the first case of its failure outside the nucleus. Nearly all the experimental work on the passage of electrons through matter reported in the last year has dealt with this problem. v. Droste[62] has carefully investigated the γ radiation emitted when the electrons from Ra E are retarded by atoms of both large and small atomic number. The results given in table 3 show the number of quanta produced per disintegration electron in three spectral ranges in which the absolute sensitivity of the counter was known.

Table 3. Retardation spectrum of Ra E in gold, lead and graphite

$h\nu$ (ekv.)	10^{-3} mm.		Pb, 0·5 mm.		C, 3·5 mm.	
	Obs.	Calc.	Obs.	Calc.	Obs.	Calc.
300	$3{\cdot}3_7 \times 10^{-5}$	$4{\cdot}97 \times 10^{-5}$	$0{\cdot}79 \times 10^{-3}$	$1{\cdot}07 \times 10^{-3}$	$0{\cdot}53 \times 10^{-4}$	$0{\cdot}69 \times 10^{-4}$
250	$5{\cdot}6_8$	7·82	1·3	1·65	0·90	1·05
200	$10{\cdot}4$	12·5	$2{\cdot}4_7$	2·99	1·65	1·91

The agreement between the observed number of quanta and those calculated from the Bethe-Heitler theory is as good as could be expected when the great experimental difficulties involved are considered.

For fast electrons ($\sim 10^9$ ev.) the atomic cross-section for radiation σ is given by the relationship

$$\sigma = \sigma_0 (4 \log 183 z^{-\frac{1}{3}} + 2/9),$$

with

$$\sigma_0 = \frac{z^2 r_0^2}{137} \quad \text{and} \quad r_0 = e^2/mc^2.$$

The rate of loss of energy $\partial E/\partial x$ of a particle of energy E passing through a plate of thickness x with n atoms per cm.^{3} is $E\sigma n$. Thus the rate of diminution of energy is proportional to the energy of the particle. The validity of this relationship has been investigated by Anderson and Neddermeyer[64] who have continued their experiments upon the energy lost by shower particles* on passing through a lead plate (0·35 cm.). The new results are given in table 4.

Table 4. Energy loss for fast electrons in lead

Energy of particle (10^6 ev.)	50	50–100	100–150	150–200	200–400
Average loss per cm. of lead:					
Observed	42	82	178	191	358
Calculated	50	110	175	248	378

* See the latter part of this section.

Anderson and Neddermeyer concluded that there is no evidence for the breakdown of the theory at these high energies. As will be seen later, many important results have been deduced from their conclusion. The opposite conclusion has been reached by Blackett and Wilson[65] and by Crussard and Leprince Ringuet[66] working with particles of higher energy (up to 5×10^9 ev.). The results are best shown in the accompanying curve taken from Blackett's paper. The ordinate

$$R = \frac{1}{E}\frac{dE}{dx} = N\sigma,$$

E being the energy of the particle passing through a plate of thickness x, should have the constant value 1·9 for lead. The steady fall in R with increasing energy

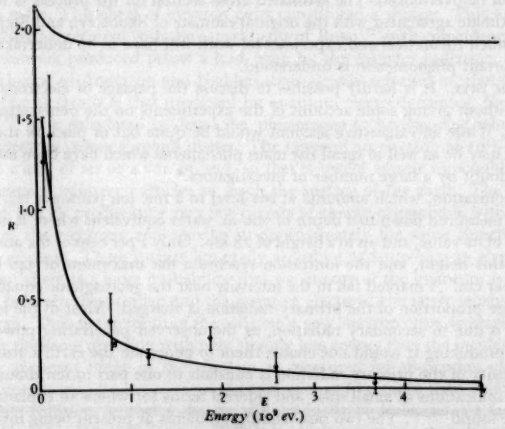


Figure 1. Energy lost by electrons in lead. ⊙ Blackett; ⊡ Crussard.

is quite incompatible with the predicted constant value of R, while the values obtained at Pikes Peak for lower energies, table 4, can be fitted on to Blackett's curve. This discrepancy will be returned to later.

Another way in which an electron can lose energy in passing through an atom is the formation of electron pairs. The theory of this process has been given by several authors[63]; the cross-section for 2×10^8 ev. electrons in lead is about 10^{-26} cm^2.

Like the cross-section for pair formation by γ rays which is some hundred times greater it depends upon the second power of the atomic number. The effect has been reported by Skobelzyn and Stephanowa[67] and Alichow and Kosalew[68] with a very much larger cross-section than 10^{-26} cm^2. Staub[69] has repeated these

experiments under conditions which avoid the presence of γ rays which would give rise to unwanted pair formation. He did not observe any electron pairs which could be attributed to incident electrons and concluded that the cross-section for this reaction was less than that for pair formation by γ rays. With the same object Champion and Barber[70] have recently examined the β-ray tracks of Ra E in a Wilson chamber containing 5 per cent of mercury dimethyl, and have twice observed the production of an electron pair. For the best track the energy of the incident electron was $1{\cdot}1 \times 10^6$ ev. and energy appeared to be conserved. Only one negative electron track was observed to proceed from the point of collision. Although this may have been due to the low energy (2000 ev.) of the other electron, the authors suggest that the possibility of its having been absorbed by the mercury nucleus should not be overlooked. The estimated cross-section for the process is 10^{-22} cm^2, in approximate agreement with the original estimate of Skobelzyn and Stephanowa. Clearly much theoretical and experimental work will have to be undertaken before this important phenomenon is understood.

Cosmic rays. It is hardly possible to discuss the passage of electrons through matter without giving some account of the experiments on the penetrating cosmic radiation. While an exhaustive account would be quite out of place in the present report, it may be as well to recall the main phenomena which have been established without doubt by a large number of investigators.*

The ionization, which amounts at sea-level to a few ion pairs per c.c. per sec., has been measured down to a depth of 700 m. water equivalent where it is reduced to 1/1000 of its value, and up to a height of 28 km. Only 2 per cent of the atmosphere is above this height, and the ionization reaches a flat maximum of 340 ion pairs per sec. per cm^3. A marked fall in the intensity near the geomagnetic equator shows that a large proportion of the primary radiation is charged. Most of the ionization observed is due to secondary radiation, as the observed penetrating power of the particles producing it would not enable them to penetrate the earth's atmosphere. The intensity of the primary radiation is constant to one part in ten thousand, but there are indications of small solar and sidereal terms for which an explanation can readily be found[73, 74]. The two outstanding problems at present being investigated are the nature of the primary radiation, and the way in which it produces and comes into equilibrium with the less penetrating secondary radiation. Some progress has been made with the latter problem, but there is a great difference of opinion on the former. Arguments have been put forward for positive and negative electrons, protons, negative protons, γ rays and neutrons.

The most important phenomenon in the production of the secondary radiation is shower-formation, in which a large number of positive and negative electrons are produced simultaneously in a small region of space. The total energy of a shower may be as high as 10^{11} ev. All the evidence suggests that this energy comes entirely from the primary particle and that the formation of the shower is frequently a tertiary and not a secondary phenomenon. If a high-pressure ionization chamber

* A much more detailed account is to be found in the Guthrie lecture[71] for 1935, and more recently in a general article by Regener[72].

is set up connected to an electrometer large numbers of ions are occasionally produced in the gas. This phenomenon is known as a burst and is usually assumed to be identical with a shower.* Nearly all workers agree that the ratio of showers to total radiation increases with height. One of the great difficulties in interpreting the observed phenomena is that there is no satisfactory experimental method of unambiguously identifying a very fast particle from its track in a Wilson chamber, while, on the other hand, it is not at all certain that the behaviour of high-energy particles can be satisfactorily predicted. Until recently a commonly held view, in spite of the obvious objections that could be brought against it, was that the primary radiation consisted of high-energy positive and negative electrons which in their passage through matter suddenly produced a large number of γ rays (with energies about 10^8 ev.) which were not very penetrating owing to the formation of electron pairs. This sheaf of electron pairs constituted a shower. This explanation is in excellent agreement with the experiments of Rossi[77] with coincidence counters on the showers produced below a lead plate by the primary radiation. Adopting the conclusion of Anderson and Neddermeyer already referred to, that the energy lost by fast electrons is that predicted by quantum electrodynamics, Bha Bha and Heitler[78] have calculated in detail the phenomena to be expected when a high-energy electron passes through matter. The range of an electron of 10^{12}-ev. energy is about 2 km. of air or 4 cm. of lead, so that it would appear impossible for even such energetic primary particles to reach the surface of the earth. The energy is, however, lost by producing a very few quanta of energy comparable with the initial energy of the electrons and moving in approximately the same direction as the original electron. These quanta are easily absorbed and produce an electron pair; moreover, there is a large probability that the energy of the quantum is unequally divided between the positive and the negative electron; the latter receives most of the energy. Thus as the result of the collision a new electron often appears travelling in nearly the same direction with only slightly less energy than the incident electron and the apparent range of the electron as measured by absorption experiments will be very large. In any case an electron cannot lose more than half its energy in a single collision, so that a shower will gradually be built up as a fast electron passes through a layer of matter. Detailed calculation shows that this shower will closely resemble a cosmic-ray shower. The main results of the calculation are reproduced in figure 2, which shows the number $\bar{N}$ of electrons of energy larger than E produced when an electron of energy E_0 passes through matter of thickness l.† The various curves $n=0$, $n=1$ refer to the initial particle, primary, secondary, etc., electrons. The curve is for a value of E such that $\log\left(\frac{E_0}{E}\right)=5$.

It will be seen that this curve reproduces the main features of the experiments of Rossi already mentioned, except for their extension to large thicknesses of lead. Clearly this theory cannot account for the most penetrating cosmic-ray particles which can produce coincidences in counters separated by 100 cm. of lead.

* Clay, 't Hooft, Dey and Wiersma[75], have recently reached the opposite conclusion.

† l is a generalized thickness proportional to the actual thickness in cm.

An entirely different point of view has been adopted by Heisenberg[79]. The equations of quantum theory represent the interaction energy of the electron and the virtual oscillators representing the light, namely, that phenomenon which was once called its reaction on the ether. The interaction energy of an electron is, however, also determined by other quantities. These interactions are in fact the basis of Fermi's theory of β radioactivity. Consideration of the two types of interaction shows a fundamental difference. The equations of electrodynamics can be expressed (if the proper mass of the electron can be neglected as it may be for high energies) in a form which contains no fundamental length. Correspondingly the wave-length of the electron does not enter into the cross-section for radiation and the cross-section (calculated by Bethe and Heitler), as will be remembered, is

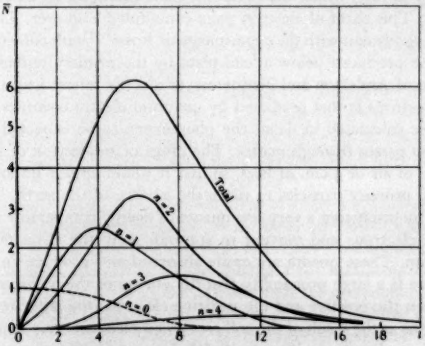


Figure 2. Shower-production by collision.

independent of the energy of the incident particle. On the contrary, the equations for the interaction of an electron which account for β radioactivity contain terms of the type

$$\int f.\psi^*_{\text{electron}}\,\alpha_\lambda\psi_{\text{neutrino}}\,\psi^*_{\text{proton}}\,\alpha_\lambda\psi_{\text{neutron}}\,dV,$$

in which f has the dimensions of cm^2, so that even when the rest mass of the electron can be neglected one may expect a cross-section which will depend upon whether or no the wave-length of the incident particle is comparable with the fundamental length $f^{\frac{1}{2}}$. Heisenberg considers the particular case of a proton colliding with another atom whose nucleus can ensure the conservation of momentum. The following changes are possible:

Proton $\rightarrow$ neutron + positron + neutrino .....(*a*),

or proton + electron + positron + 2 neutrinos .....(*b*),

or proton + n electrons + n positrons + $2n$ neutrinos .....(*c*).

The cross-sections σ for the various processes are calculated to be

(1) $$\sigma_1=\left(\frac{Ze^2}{\hbar c}\right)^2 f^2k^2,$$

(2) $$\sigma_2=\left(\frac{Ze^2}{\hbar c}\right)^2 f^4k^6,$$

(3) $$\sigma_3=\left(\frac{Ze^2}{\hbar c}\right)^2 f^{4n}k^{8n-2},$$

where $\lambda=2\pi/k$ is the mean wave-length of the particles. Thus if $k\sim 1/of$ the production of a shower of electrons in one process becomes probable. This is a fundamental difference from the radiation theory in which the production of n quanta is $(137)^n$ times less probable than the formation of a single quantum. If the value of the fundamental length $f^{\frac{1}{2}}$ is calculated from Fermi's original theory (where $f=$ Fermi's $g/\hbar c$), the energy of a shower-producing particle is much too large, but reasonable values are obtained by adopting the later theories in which the interaction energy depends upon the derivatives of the neutrino and wave functions. The neutrinos produced in this reaction will have as much energy as the electrons, and it is possible to calculate their absorption cross-section. This is about 10^{-26} cm^2, and Heisenberg suggests that they may constitute the non-ionizing link between showers. Some recent experiments indicate that both types of shower-producing process are taking place. The number of showers observed beneath a lead plate increase quadratically with the thickness of the plate when they are detected as coincidences between three counters[80]. As more counters* are used to detect the shower the number of showers approaches to a linear dependence upon the thickness of lead[81]. The former result corresponds to the gradual building up of the shower as envisaged by Bha Bha and Heitler, while the latter corresponds to the sudden production of all the shower particles in one collision as imagined by Heisenberg.

Returning now to the passage of the shower electrons (10^8 ev.) through a lead plate observed by Anderson and Neddermeyer. Reference to the table on p. 184 shows that the electron loses about half its energy on passing through the lead plate (0·35 cm.), and therefore very few particles should penetrate a 2-cm. plate. Accordingly these authors replaced their lead plate by a platinum plate 1 cm. thick (equivalent to 1·96 cm. of lead or 1·86 cm. for a Z^2 absorption law). They found[83] that the particles entering the plate could be divided into two types, the one which lost nearly all its energy while the other suffered but a small diminution of energy on traversing the plate. Moreover, the particles with the large energy loss entered the chamber as members of a shower while the penetrating particles entered singly. The tentative conclusion drawn was that there may be negatively charged particles present in cosmic rays which are not electrons. Similar conclusions have been reached by Street and Stevenson[84], who observed particles which had passed successively through a pair of counters separated by 10 cm. of lead, a Wilson chamber,

* This is not in agreement with the results of Hu Chien Shan[83].

a pair of counters, 3 cm. of lead and a second Wilson chamber with three separate lead plates 1 cm. thick in it. The observed penetrations are given in table 5.

Table 5

$H\rho \times 10^{-6}$	% total tracks	Fraction with range greater than 3 cm. of lead	Fraction with range greater than 6 cm. of lead
>5	49	1	1
2·5 to 5	19	0·9	0·7
1·5 to 2·5	23	0·9	0·7
0·7 to 1·5	9	0·7	0·3

The authors could distinguish from the appearance of the track protons for $H\rho < 2{\cdot}5 \times 10^6$, so that the penetrating particles cannot have been negative protons. It will be observed that the validity of these conclusions depends upon the validity of the quantum theory of the passage of electrons through matter. Most of the above measurements refer to particles whose energy is in the region where Blackett has observed an increasing deviation from the theoretical absorption curve, so that it is clear that great caution must be used in interpreting the results obtained.

Finally attention must be drawn to two long papers by Blackett[85] in which the energy-range relationship is extended to very fast particles (4×10^{10} ev.) by combining his own measurements on the cosmic-ray energy spectrum with absorption curves measured by other workers in the manner first used by Street, Woodward and Stevenson[86].

(*d*) *Protons.* In spite of the large number of fast proton sources in existence very little quantitative work has been carried out on the general phenomena occurring when protons pass through matter. A careful investigation of the scattering of protons by hydrogen has been carried out by Tuve, Heydenburg and Hafsted[87]. A narrow beam of protons of known energy and separated from other ions by magnetic analysis was allowed to enter a cylindrical scattering chamber filled with pure hydrogen at 12 mm. pressure, through a thin aluminium window (S.P. $= = \frac{1}{2}$ cm.). The protons scattered from a small central volume were selected by a series of diaphragms and detected by means of an ionization chamber and linear amplifier. The ionization chamber and slit system could be rotated about the axis of the scattering chamber so as to select particles scattered at any required angle. The constancy of the primary proton beam whose magnitude was about 5×10^{-9} amp. was ascertained by collecting the beam in a Faraday cylinder after it had traversed the evacuated scattering chamber. Obviously the ionization of the scattering gas prevented this measurement being made during an experiment, so that measurements were made upon the beam immediately before and after each run. Results in which the intensity of the beam was found to have changed were discarded. In the final measurements a second ionization chamber orientated at a fixed angle to the beam was placed in the scattering chamber. Its use enabled any large variation in the proton stream to be detected immediately. The results obtained at 600, 700, 800 and 900 kv. compared with the theoretical scattering due to an

inverse square-law repulsion are shown in the diagram. In some later experiments using a point counter to detect the protons, measurements were also carried out down to 200 kv. Owing to the identity of the projected and scattered particles the theoretical curve is different from the Rutherford scattering curve even for repulsion obeying Coulomb's law[87a]. This has been taken into account in drawing the diagram, so that the observed departure from the theoretical curve is entirely due to the breakdown of Coulomb's inverse square-law for the force between the protons. The anomaly can be accounted for by a phase shift of the first partial wave function. The proton proton interaction obtained in this way is nearly the same as that between a proton and a neutron in a ^1S state, as obtained from experiments on the scattering of neutrons in hydrogenous substances.

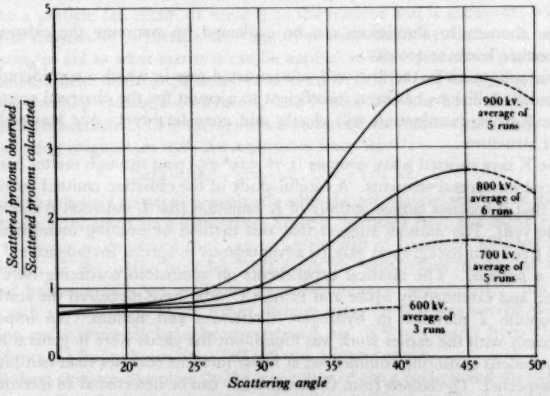


Figure 3. Scattering of protons in hydrogen.

The anomalous scattering of protons in other light elements has been investigated by Dymond[88], who has found somewhat unexpected results. A stream of 200 kv. protons passed, after magnetic analysis, into a scattering chamber in the centre of which was mounted a thin scattering foil which covered the entrance to a completely enclosed Faraday cylinder. The metal films (which were prepared by a special technique to avoid the presence of tungsten, a frequent impurity in films prepared by evaporation from a tungsten spiral) were sufficiently thin ($3{\cdot}7 \times 10^{-7}$ cm.) to allow most of the proton beam to traverse them unscattered so that the protons scattered backwards by the foil could safely be assumed to have had only one nuclear collision. The recoil protons were too few to be measured by collection in a Faraday cage, so they were counted by measuring the combined secondary current produced when they fell upon a platinum electrode by means of an electrometer valve and bridge circuit. The scattering in the light elements (beryllium, boron and

aluminium) was compared with that in silver which was assumed to scatter according to the Rutherford law. Boron was found to scatter normally but aluminium and beryllium showed a definite anomaly.

Table 6

Element	Excess scattering observed (per cent)	
	At 110°	At 150°
Be	13·5	4·4
B	0	0
Al	9·7	17·6

The anomaly in aluminium can be explained by assuming the existence of D resonance levels at 200 ekv.

This appears to be the first well-substantiated case in which a consideration of head on (S) collisions has been insufficient to account for the observed scattering, and its further examination will clearly add considerably to our knowledge of nuclear structure.

The X rays emitted when protons ($1{\cdot}76 \times 10^6$ ev.) pass through matter has been measured for several elements. A careful study of the radiation emitted by zinc[89] shows that it consists almost entirely of *K* radiation (the *L* radiation is too soft to be observed). The authors suggest that this method of exciting monochromatic X rays free from background may be advantageous in special investigations.

(*e*) α *particles*. The classical experiments on anomalous scattering have been repeated and extended by Mohr and Pringle[90] who have measured the scattering of polonium α particles in hydrogen, deuterium and helium. No important discrepancy with the earlier work was found, but the yields were in general higher and anomalous scattering commenced at lower incident energies than had hitherto been suspected. Deviations from Coulomb's law can be detected at an internuclear distance as great as 10^{-12} cm.; the occurrence of anomalous scattering for α particles of 1-cm. range may well account for the remarkably different values for the absolute scattering cross-section obtained by different workers in this field. The experimental data obtained was analysed in a novel way, and the authors were able to show that their results could consistently be interpreted in terms of a Gamow potential barrier if, and only if, more than one phase constant were taken into account.

The production of induced radioactivity by α-particle bombardment of magnesium and aluminium has been examined by Chang[91] and his coworkers, who find that resonance effects are much more marked when a proton is emitted as in the reaction

$$^{25}_{12}\mathrm{Mg} + {}^{4}_{2}\mathrm{He} \rightarrow {}^{28}_{13}\mathrm{Al} + {}^{1}_{1}\mathrm{H}$$

than in the alternative reaction in which a neutron is emitted.

It is therefore suggested that the phenomenon of resonance is due to exceptional transparency of the potential barrier to the emitted, not to the incident particle.

A clear-cut example which may be explained in this way has been found in the disintegration of nitrogen by fast neutrons[92]:

$$^{14}_{7}\mathrm{N} + ^{1}_{0}n \rightarrow ^{11}_{5}\mathrm{B} + ^{4}_{2}\mathrm{He}.$$

When this reaction is examined with an ionization chamber and linear amplifier, it is found that four groups of α particles are emitted with energies 1·4, 2·0, 2·5 and $3{\cdot}1 \times 10^{6}$ ev. respectively. This is in all probability due to a potential barrier of the intermediate nucleus formed by the incident neutron being particularly transparent to α particles of those energies. The alternative explanation that the discrete groups are due to resonance entry of the neutron cannot be excluded but appears to be unlikely.

This interpretation of resonance disintegration avoids the difficulty of assuming that the α particle can retain its identity in the nucleus and is consonant with the modified theories of α-particle disintegration introduced by Bethe. It will be interesting to see to what extent it can be applied to other resonance phenomena.

Bombardment with α particles is the most fruitful method of producing positron-emitting radioactive elements, and a new one has been discovered by Ellis and Henderson[93] in a careful examination of the radio elements produced by irradiating magnesium with the α particles from Ra C′:

$$^{24}_{12}\mathrm{Mg} = ^{4}_{2}\mathrm{He} \rightarrow ^{27}_{14}\mathrm{Si} + ^{1}_{0}n^{1}; \quad ^{27}_{14}\mathrm{Si} \rightarrow ^{27}_{13}\mathrm{Al} + e^{+}.$$

Even in this field a start is being made with the production and use of artificial α particles accelerated in a cyclotron. Since α particles have the same ratio of mass to charge as deuterons they can be accelerated in cyclotron adjusted for the production of the latter particles. Although no satisfactory source of doubly charged helium ions is yet available, some preliminary experiments[94] have been made with 5×10^{-8} amp. of 7×10^{6}-ev. helium ions. In this way several positron-emitting radio elements have been produced, i.e. K^{38} (7·5 min.) from chlorine, Fe^{55} (8·9 min.) from chromium, Ga^{68} (68 min.) from copper.

All these changes take place by the absorption of the α particle and the expulsion of a neutron, but doubtless other types of reaction will be found as the experiments are extended. There is little doubt that the use of artificial α particles will prove to be a valuable method of extending our knowledge of radioactive isotopes.

References

(48) Rusinow and Sagiadak. *Phys. Z. Sowjet.* **10**, 203 (1937).
(49) (*a*) Lea. *Proc. roy. Soc.* A, **150**, 637 (1935).
(*b*) Collie and Griffiths. *Proc. roy. Soc.* A, **155**, 434 (1936).
(50) Kikuchi. *Nature, Lond.*, **137**, 398 (1930); *Proc. phys.-math. Soc. Japan*, **18**, 115 (1936).
(51) Aoki. *Nature, Lond.*, **139**, 372 (1937).
(52) Grahame, Seaborg and Gibson. *Phys. Rev.* **51**, 590 (1937).
(53) Gibson, Seaborg and Grahame. *Phys. Rev.* **51**, 370 (1937).
(54) (*a*) Tuve and Hafsted. *Phys. Rev.* **50**, 490 (1936).
(*b*) Booth and Hurst. *Nature, Lond.*, **138**, 1011 (1936). *Proc. roy. Soc.* A, **161**, 248 (1937).
(*c*) Dunning, Pegram, Fink and Mitchell. *Phys. Rev.* **48**, 265 (1935).
(55) Bethe and Bache. *Rev. Mod. Phys.* **8**, 117 (1936).
(56) Morse Fisk and Schiff. *Phys. Rev.* **51**, 706 (1937).

(57) Harkins, Karmen Newson and Gans. *Phys. Rev.* **50**, 980 (1936).
(58) Pool, Cork and Thornton. *Phys. Rev.* **52**, 41 (1937).
(59) Heyn. *Nature, Lond.*, **138**, 723 (1936); *Physica*, **4**, 160 (1937).
(60) (*a*) Pool, Cork and Thornton. *Phys. Rev.* **51**, 890 (1937).
(*b*) Chang, Goldhaber and Sagane. *Nature, Lond.*, **139**, 962 (1937).
(61) Walke. *Phys. Rev.* **51**, 439 (1937).
(62) Frhr v. Droste. *Z. Phys.* **104**, 335 (1937).
(63) Heitler. *Quantum Theory of Radiation* (1936). Oxford: Clarendon Press.
(64) Anderson and Neddermeyer. *Phys. Rev.* **50**, 263 (1936).
(65) Blackett and Wilson. *Proc. roy. Soc.* A, **160**, 304 (1937).
(66) Crussard and Leprince Ringuet. *C.R. Acad. Sci., Paris*, **204**, 240 (1937).
(67) Skobelzyn and Stephanowa. *J. de Phys.* **8**, 61 (1935).
(68) Alichow and Kosalew. *Z. Phys.* **90**, 249 (1934).
(69) Staub. *Helv. phys. Acta*, **9**, 306 (1936).
(70) Champion and Barber. *Nature, Lond.*, **140**, 105, (1937).
(71) Compton. Guthrie Lecture, *Proc. phys. Soc.* **47**, 747 (1935).
(72) Regener. *Naturwissenschaften*, **25**, 1 (1937).
(73) Compton and Getting. *Phys. Rev.* **47**, 817 (1935).
(74) Vallarta. *Nature, Lond.*, **139**, 839 (1937).
(75) Clay, 't Hooft, Dey and Wiersma. *Physica*, **4**, 121 (1937).
(76) Leprince Ringuet and Crussort. *J. Physic.* **204**, 112 (1937).
(77) Rossi. *Z. Phys.* **68**, 64 (1931).
(78) Bhabha and Heitler. *Proc. roy. Soc.* A, 159, 432 (1937).
(79) Heisenberg. *Z. Phys.* **101**, 533 (1936).
(80) Morgan and Nielsen. *Phys. Rev.* **50**, 882 (1936).
(81) Watase. *Nature, Lond.*, **139**, 672 (1937).
(82) Hu Chien Shan. *Proc. roy. Soc.* A, **158**, 588 (1936).
(83) Anderson and Neddermeyer. *Phys. Rev.* **51**, 884 (1937).
(84) Street and Stevenson. *Phys. Rev.* **51**, 1005 (1937).
(85) Blackett. *Proc. roy. Soc.* A, **159**, 1, 19 (1937).
(86) Street, Woodward and Stevenson. *Phys. Rev.* **47**, 891 (1935).
(87) Tuve, Heydenburg and Hafsted. *Phys. Rev.* **50**, 806 (1936).
(87 *a*) Mott and Massey. *Theory of Atomic Collisions*. Oxford: Clarendon Press.
(88) Dymond. *Proc. roy. Soc.* A, **157**, 302 (1936).
(89) Livingstone, Genevese and Konopinski. *Phys. Rev.* **51**, 802 (1937).
(90) Mohr and Pringle. *Proc. roy. Soc.* A, **160**, 190 (1937).
(91) (*a*) Waring and Chang. *Proc. roy. Soc.* A, **157**, 652 (1936).
(*b*) Chang and Szalay. *Proc. roy. Soc.* A, **159**, 72 (1937).
(92) Wilhelmy. *Naturwissenschaften*, **25**, 173 (1937).
(93) Ellis and Henderson. *Proc. roy. Soc.* A, **156**, 358 (1936).
(94) (*a*) Henderson, Ridenour, White and Henderson. *Phys. Rev.* **51**, 1102 (1937).
(*b*) Henderson and Ridenour. *Phys. Rev.* **52**, 40 (1937).

§ 5. β RADIOACTIVITY

No generally accepted theory of β radioactivity has yet been put forward, and the numerous artificially prepared β-radioactive substances have been intensively studied especially in the United States, where relatively intense sources can be prepared by means of cyclotrons.

However, it is generally agreed* that the end-point of the continuous β-ray spectrum represents the energy change of the transformation, and this important physical constant has been measured for a large number of substances in the period

* See last year's *Report*.

under review. The results obtained by numerous workers for artificially radioactive substances have been collected and are given below. Using these results Mitchell[95] has shown that all the artificially radioactive substances lie upon one of three Sargent curves, and has tried to deduce the spin quantum number of the nuclei concerned from them.

Table 7. End-points of β-ray spectra; positron emitters

Active element	Half life	End-point ($10^6 \times$ ev.)
^{11}C	21·0 min.	1·3
^{13}N	10·3 min.	1·45
^{15}O	126 sec.	2·0
^{17}F	1·16 min.	2·4
^{34}Cl	33 min.	3·0
^{64}Cu	12·8 hr.	0·79

Table 8. End-points of β-ray spectra; electron emitters

Active element	Half-life	End-point ($10^6 \times$ ev.)
^{8}Li	0·88 sec.	11·2 (10·0 obs.)
^{12}B	0·02 sec.	13·0 (11·5 obs.)
^{20}F	12 sec.	5·9
^{24}Na	15·5 hr.	1·95
^{31}Si	145 min.	2·05
^{32}P	14·5 d.	2·15 (1·69 also 1·95)
Cl	40 min.	1·5, 6·1
^{41}A	108 min.	1·5, 5
^{42}K	12·2 hr.	1·4, 4·4
^{56}Mn	2·5 hr.	1·2, 2·9 (3·2)
^{64}Cu	12·8 hr.	0·83
^{76}As	26 hr.	1·09, 3·4
^{80}Br	18 min.	(2·00)
^{80}Br	4·2 hr.	(2·05)
^{82}Br	36 hr.	(0·85)
Rh	44 sec.	2·8 (2·5)
Ag	22 sec.	2·8 (2·7)
In	13 sec.	3·2
In	54 min.	1·45
I	25 min.	(2·1)
Dy	2·5 hr.	1·4
Ir	19 hr.	(2·2)
Au	2·7 d.	(1·15)

All end-points were obtained by using the Konopinski-Uhlenbeck extrapolation formulae except those in brackets, which are directly observed end-points.

The determination of the end-point of the spectrum is experimentally difficult, as it is not known how the energy-distribution curve approaches the axis. The most satisfactory method is to make a large number of measurements on the electron tracks in a Wilson chamber; this method is laborious and is limited to strong sources. If great care is taken to avoid scattering satisfactory measurements may be made with the conventional β-ray spectrograph; the electrons should be detected with a coincidence counter which avoids the large natural background of a single counter. It is not surprising that attempts have been made to find a reliable method

of extrapolating to the end-point from measurements of the much more numerous electrons of lower energy. According to Fermi's theory the shape of the continuous spectrum is given by the relationship

$$(1+\eta^2)^{\frac{1}{2}} = C - K\left[\frac{N}{f(\eta Z)}\right]^{\frac{1}{2}}.$$

N is the number of disintegration electrons which have momentum ρ expressed in units of mc or $H\rho/1700$ gauss-cm. $C=(1+\eta_0^2)^{\frac{1}{2}}$, where η_0 is the end-point of the spectrum.

The $f(\eta Z)$ can be evaluated from the theory, for each atomic number Z. Thus if the experimental values of $(1+\eta^2)^{\frac{1}{2}}$ are plotted against $[N/f(\eta Z)]^{\frac{1}{2}}$ one should obtain a straight line which can be extrapolated to cut the axis at C; hence the required quantity η_0 can readily be found. An examination of the β-ray spectra of radio silicon and radio sodium[96] showed that an extrapolation formula derived from Konopinski and Uhlenbeck's[97] modification of Fermi's theory was more satisfactory. This method of obtaining end-points has been generally adopted by American workers, and its validity is a matter of some practical importance. It may be noted in passing that according to the modified theory the β-ray distribution curve has a fourth-order contact with the energy axis so that a direct determination of the end-point would be difficult. With the object of testing this method Langer and Whittaker[98] measured the β-ray spectrum of a strong source of Ra E, using a β-ray spectrograph specially arranged to avoid scattering and detecting the electrons with a coincidence counter. The end-point was observed at $H\rho=5330\pm70$ ($1{\cdot}17.10^6$ ev.), while the end-point obtained by extrapolating the straight part of a Konopinski-Uhlenbeck plot was at $H\rho=6250$. Lyman[99], who carried out a similar experiment using a single counter, found the end-point at $H\rho=5280$ and the Konopinski-Uhlenbeck end-point was at $H\rho=7150$. It may, however, be objected[100] that the Ra E disintegration is a forbidden one, so that the method may be reliable after all for allowed transitions.

Thus, in spite of the careful measurements which have been made, the value of this method of extrapolation is still uncertain.

The problem may be attacked in a completely different way by examining the radiations emitted by a positron radioactive substance. According to theory[101] there is quite a large possibility that such a nucleus will reach its stable configuration by absorbing an electron from the K shell instead of emitting a positron.

Experiments have been carried out by Jacobsen with the positron emitter radio scandium ($^{45}_{21}$Sc). He observed its radiations in a Wilson chamber but could find no short electron tracks which he could attribute to Ca K radiation; he estimated that the probability of forming Ca by electron absorption was less than $\frac{1}{100}$. The probability to be expected on theoretical grounds depends markedly upon the type of neutrino interaction which is adopted; Konopinski-Uhlenbeck interaction gives so high a probability that Jacobsen's experiment may be said to have excluded it. Hoyle[103] has investigated the various ways in which the original Fermi theory can be modified and has found another variation of the form of the interaction energy which reduces the probability of electron absorption to $\frac{1}{500}$. As the shape of the

β-ray spectrum, however, also depends upon the type of interaction adopted, it may be possible to test Hoyle's hypothesis experimentally; he predicts that a positron-emitting substance will emit a large proportion of the disintegration positrons with very little energy. Now that medium atomic number positron emitters can be prepared with a cyclotron, this prediction will doubtless soon be put to the test of experiment.

It not infrequently happens that the nucleus is left in an excited state after the emission of the β particle; the daughter nucleus subsequently returns to its ground state by emitting a γ ray. The relationship between the β-ray spectrum and the energy of the subsequently emitted γ ray was discussed some time ago by Ellis and Mott[104], and the hypothesis they put forward has been tested experimentally by Bothe and Maier-Leibnitz[105] in an ingenious series of experiments. Working with Ra C, they registered the coincidences between a γ-ray counter placed near the source and an electron counter so situated that it only received electrons of a particular energy. They obtained the results shown in table 9.

Table 9

Electron energy (ev. $\times 10^{-6}$)	0·21	0·47	1·09	1·86
Coincidences per 1000 particles	1·57	1·52	0·83	0·02

The energy spectrum of Ra C extends to $3{\cdot}12 \times 10^6$ ev., but no coincidences were observed above 2×10^6 ev.

These results are not in agreement with the hypothesis that Ra C′ is hardly ever formed in its ground state as assumed by Ellis and Mott, and suggest that the daughter nucleus is formed either in the ground state or in one excited state only.

References

(95) Mitchell. *Phys. Rev.* **52**, 1 (1937).
(96) Kurie Richards and Paxton. *Phys. Rev.* **48**, 167 (1935).
(97) Konopinski and Uhlenbeck. *Phys. Rev.* **48**, 7 (1935).
(98) Langer and Whittaker. *Phys. Rev.* **51**, 713 (1937).
(99) Lyman. *Phys. Rev.* **51**, 1 (1934).
(100) Lamb and Pollard. *Phys. Rev.* **51**, 682 (1937).
(101) (*a*) Moller. *Phys. Rev.* **51**, 84 (1937); *Phys. Z. Sowjet.* **11**, 9 (1937).
(*b*) Mercier. *Nature, Lond.*, **139**, 798 (1937); *C.R. Acad. Sci., Paris*, **204**, 1117 (1937).
(*c*) Fietz. *Z. Phys.* **104**, 553 (1937).
(102) Jacobsen. *Nature, Lond.*, **139**, 879 (1937).
(103) Hoyle. *Proc. Camb. phil. Soc.* **33**, 286 (1937); *Nature, Lond.*, **140**, 235 (1937).
(104) Ellis and Mott. *Proc. roy. Soc.* **141**, 502 (1933).
(105) Bothe and Maier-Leibnitz. *Z. Phys.* **104**, 604 (1937).

SLOW NEUTRONS

By P. B. MOON, M.A., Ph.D.

§1. SUMMARY OF EARLIER WORK

The properties of slow neutrons have been mentioned only briefly in earlier *Reports*, and a separate article on this subject was foreshadowed last year. It therefore seems desirable to give a summary of work published prior to about the end of 1935, in order to provide a proper background to the discussion of more recent developments.

In 1934 Fermi and his associates[1] discovered the production of β-radioactive nuclei when many elements are irradiated by neutrons from a radon-beryllium source. Typical reactions of this type are

$$\begin{matrix} {}^{19}_{9}\mathrm{F} + {}^{1}_{0}n \rightarrow {}^{16}_{7}\mathrm{N} + {}^{4}_{2}\mathrm{He} \\ \qquad\qquad\quad \searrow {}^{16}_{8}\mathrm{O} + \beta \end{matrix} \qquad \ldots\ldots(a),$$

$$^{127}_{53}\mathrm{I} + {}^{1}_{0}n \rightarrow {}^{128}_{53}\mathrm{I} \rightarrow {}^{128}_{54}\mathrm{Xe} + \beta \qquad \ldots\ldots(b).$$

The collision cross-sections for such processes are of the order of 10^{-24} cm^2, which is comparable with the size of a nucleus as found by experiments on the scattering of (e.g.) α particles.

Shortly afterwards Fermi and his associates[2] found, on surrounding the source and specimen with water or paraffin wax, that neutrons could be made slow as a result of collisions with protons, and that simple capture processes, such as equation (*b*), occur much more readily with slow neutrons. They suggested that many of the neutrons reached thermal equilibrium with the wax before being captured by a carbon or a hydrogen nucleus; this was confirmed by the experiments of Moon and Tillman[3], who found a marked influence of the temperature of the wax upon the capture of the slow neutrons by various elements.

A volume of wax or water near a source of fast neutrons is therefore traversed by a flux of neutrons whose energies range from millions of volts to a small fraction of a volt, and any division into fast and slow must be partly arbitrary. It happens that knock-out reactions, such as (*a*), require the neutron to have a kinetic energy of the order of a million volts, while capture processes, such as (*b*), reach their greatest efficiency at energies below 100 v.* A tentative dividing line may thus safely be drawn in the intermediate region of energy—say at 50,000 v.

Fermi and his collaborators[2] measured the absorption of slow neutrons by various elements, using the activity induced in rhodium as a measure of the incident and transmitted flux of neutrons. They found that strong absorption is not confined to elements which show strong α-ray activity after irradiation; indeed, the highest absorption coefficients were found for boron, lithium, cadmium and mercury,

* Although capture processes were first discovered with supposedly fast neutrons, this was almost certainly due to a small admixture of slow neutrons produced by collisions in the source, the specimen, or neighbouring bodies.

which show no appreciable activity. The absorption by boron and lithium proved to be due to the capture of a neutron by ^{10}B and ^{6}Li with the immediate emission of an α-particle[4,5], while for cadmium and mercury, which possess many consecutive stable isotopes, the capture of the neutron yields a stable nucleus in an excited state, and γ-rays only are emitted. Nitrogen has since been found to capture slow neutrons with the emission of a proton[6,7], but it is only for these three light elements (Li, B, N) that the capture of a *slow* neutron causes a heavy particle to leave the capturing nucleus ***immediately***; in all other instances the slow neutron is simply captured with the formation of an isotopic nucleus. This nucleus may or may not be stable, but this is of no matter so far as the initial nuclear reaction is concerned, important though it may be to the experimenter.

An ionization chamber lined with lithium or boron, or filled with BF_3 gas, is a very sensitive detector of slow neutrons; the chamber is coupled to a valve amplifier and the individual α particles recorded. In comparison with the use of a Geiger-Müller counter to measure the radioactivity induced in (e.g.) rhodium, this way of counting neutrons has the advantages of reliability, greater sensitivity and lower natural count.

Using this method, Dunning and collaborators[8] measured the cross-sections for collisions of slow neutrons with nearly all the elements; the geometrical conditions were well defined and such that the sum of the scattering and absorption cross-sections was measured. Separate experiments were made to find the proportion which is due to scattering, and it was found that the very large cross-sections are due almost entirely to absorption. For example, cadmium scatters far less than does copper, which has a moderate total cross-section.

The fraction of neutrons which a given absorber appears to capture was early found[9,3] to depend upon the method used to detect the neutrons, being greatest when the detector is the same element as the absorber. This means that certain elements absorb selectively neutrons of certain velocities; that is, the law connecting absorption cross-section with velocity is different for different elements.

An attempt to explain theoretically the very large capture cross-sections and their random occurrence among the elements was outlined by Fermi[2] and worked out in detail by Bethe[10] and others. In this theory the interaction of nucleus and neutron is considered as a two-body problem in quantum mechanics, the neutron being supposed to move in the field of the nucleus, which is idealized to a potential hole. If the wave-length of the neutron when inside the nucleus happens to be correctly attuned to the dimensions of the hole, resonance will occur and the chance of capture will be very high. Since the radii of heavy nuclei are of the order of 10^{-12} cm., and the wave-length of a million-volt neutron is also about 10^{-12} cm., it is reasonable that such resonances should occur; but only a few resonances (the lowest orders) can be effective, for to reduce the wave-lengths even ten times would require the potential hole to have a depth of about 10^{8} v. Since the resonances are so few they must be thousands of volts broad if the frequent occurrence of very large cross-sections for slow neutrons is to be explained. Over a region of energy small compared with the width of a nuclear resonance, the theory

leads inescapably to the conclusion that the absorption cross-section should vary inversely with the velocity of collision; this l/v law is in conflict with the experimental facts of selective absorption. Further, the theory predicts that a large absorption cross-section should always be accompanied by a large cross-section for scattering. This, again, is in contradiction with experiment.

The position towards the end of 1935 may therefore be summarized as follows: (1) Atomic nuclei capture slow neutrons either (*a*) by simple addition with the formation of an isotope of the original nucleus, which may or may not be radioactive, or (*b*) with the immediate emission of a heavy particle, as in the case of B, Li, N. (2) The absorption cross-sections vary greatly and in an apparently random way from one element to another and are frequently enormously greater (approaching 10^{-20} cm^2) than the nuclear dimensions. The dependence of absorption cross-section upon velocity is different for different nuclei. (3) Large absorption cross-sections are not accompanied by large scattering cross-sections. (4) The experimental facts cannot be explained by a two-body collision theory in which the nucleus is idealized to a potential hole.

§2. THE MANY-BODY NUCLEUS

Bohr[11, 12, 13] was the first to point out that the two-body theory of collisions of neutrons with heavy nuclei should be replaced by a many-body treatment in which account is taken of the exchange of energy between the many distinct particles which are packed into the nucleus. The basic fact is that the forces between nuclear particles are short-range forces, as opposed to the long-range (inverse square) electrostatic forces which dominate extra-nuclear phenomena. The distances between neighbouring particles are of the same order as the range of the forces, and the forces between two neighbouring particles are quite comparable with the forces between one of them and the rest of the nucleus; the nucleus may therefore be likened to a drop of liquid which is held together by the cohesive forces between neighbouring molecules. Two important consequences follow. First, a particle which strikes a heavy nucleus cannot pass through, as a fast electron passes through the extra-nuclear structure of an atom, but must lose energy by collisions with the closely packed nuclear particles. Secondly, there will be no such great differences between the binding energies of the various nuclear particles as there are between the binding energies of the inner and outer electrons of a heavy atom, and if energy is given to the nucleus it is most likely to be shared among several particles and not to go to only one or two; the energy levels of the nucleus as a whole will therefore be very numerous and closely spaced. The heavier the nucleus the more complex will be its level system and the closer the spacing.

If the incident particle is a slow neutron it will gain some kinetic energy by the attraction of the outer nuclear particles, but when it arrives inside the nucleus it will share this energy by collisions with other particles, and will not have enough energy to escape against the surface forces. In the course of time, however, sufficient energy may again become concentrated in a single neutron to enable it to escape, and we shall say that the original neutron has been scattered, although the compound nucleus has lasted for a comparatively long time. There are two ways in which the re-emission of a neutron may be forestalled; first, another particle (an α particle

or a proton) may require less energy to escape than does a neutron; and secondly, the energy may be dissipated by γ radiation. If the original neutron was slow, nearly all the energy of excitation of the compound nucleus must be concentrated on it if it is to escape; unless a charged particle is emitted γ radiation will have time to occur and the neutron will be permanently captured. Thus in any case the chance of re-emission (scattering) of a slow neutron is small.

The above discussion is entirely corpuscular and can give only a general picture of the process. The incident neutron, and indeed the whole system, has wave properties, and we must expect resonance phenomena. If the kinetic energy of the incident neutron is such that the total energy of the compound nucleus would coincide with one of its energy levels, the absorption cross-section will be exceptionally large. There will thus be resonance absorption bands whose widths* and spacing will be the same as the widths and spacing of the energy levels of the compound nucleus.

Just as the natural width of an electronic energy level is connected with its mean lifetime (i.e. with the probability of emission of radiation), so the width of a nuclear energy level will be the sum of the partial widths which represent the probabilities of emission of various particles. For heavy nuclei whose life after capturing a neutron is so long that there is time for the emission of a γ ray, the resonances will be very sharp, while for boron and lithium, which after capturing a neutron emit an α particle before a γ ray can escape, the levels will be much broader.

Independently of Bohr, Breit and Wigner[14] suggested that the two-body treatment should be abandoned, and made calculations in which the entry of the neutron was supposed to give rise to a compound nucleus in an excited state, the excess energy being carried away either by a re-emitted neutron, or by a charged particle, or by γ radiation. They considered only *one* nuclear resonance, but this is little disadvantage because in most instances the resonance level which lies nearest to the actual energy of the compound nucleus will outweigh all others.

If the neutron width Γ_n is small compared with the sum of all the other partial widths, Γ (i.e. if the probability of re-emission of a neutron is small compared with that of γ radiation for a heavy nucleus or of charged-particle emission for such nuclei as boron and lithium, and if the neutron goes first into an S state then Breit and Wigner's expression for the capture cross-section for a neutron of kinetic energy E may be written

$$\sigma_{cap} = \frac{h^2}{8\pi M (EE_r)^{\frac{1}{2}}} \cdot \frac{\Gamma_n \Gamma}{(E-E_r)^2 + \frac{1}{4}\Gamma^2} \qquad \text{.....(1),}$$

while the scattering cross-section is

$$\sigma_{sc} = \frac{h^2}{8\pi M E_r} \cdot \frac{\Gamma_n^2}{(E-E_r)^2 + \frac{1}{4}\Gamma^2} \qquad \text{.....(2).}$$

M is the neutron mass and E_r the resonance energy.

* The width will be defined as the range of *energy* over which the absorption cross-section exceeds one-half its maximum value.

The shape of the resonance is thus the same as that of the natural width of a spectral line.

At exact resonance, when $E=E_r$, we have

$$\sigma_{cap}=\frac{h^2}{2\pi ME_r}\cdot\frac{\Gamma_n}{\Gamma} \qquad \text{.....(3),}$$

$$\sigma_{sc}=\frac{h^2}{2\pi ME_r}\left(\frac{\Gamma_n}{\Gamma}\right)^2 \qquad \text{.....(4).}$$

It will be noticed that the expression for the *capture* cross-section contains not only the resonance factor but also a factor $E^{-\frac{1}{2}}$. Thus the resonances are superimposed upon a general $1/v$ variation of absorption cross-section with velocity, and for energies small compared with either the resonance energy E_r or the resonance width Γ (or both) the $1/v$ factor will be dominant. For boron and lithium, which emit α particles on capturing a slow neutron, the resonance is so broad that the $1/v$ law should hold up to some thousands of volts, but for the heavy nuclei with their closely spaced and narrow levels the resonances will be the dominant factor at energies of the order of a few volts or sometimes below a volt. Breit and Wigner's calculations have been generalized by Bethe and Placzek[15] to take account of multiple energy levels.

The many-body theory is therefore able to overcome the two great difficulties which were encountered by the one-body theory, for it interprets the phenomena of selective absorption in terms of the sharp and closely spaced energy levels of heavy nuclei, and it explains why strong absorption is not accompanied by comparably strong scattering.

§3. COLLISIONS OF NEUTRONS WITH PROTONS

Neutron-proton collisions are necessarily two-body phenomena and their theory is unaffected by the developments discussed in the last section. The theory of the scattering of neutrons by free protons, discussed in vol. 2 of these *Reports*[16], has been extended by Fermi[17] to scattering by the bound protons of paraffin wax (the most widely used experimental material) and to the capture of neutrons by protons.

If the energy of the neutron is well above the binding energy of the protons in the molecule of paraffin—say, above 1 ev.—the proton may be treated as free and stationary and the scattering cross-section is given by the expression

$$\sigma_{sc}=\frac{h^2}{\pi M}\left[\frac{3}{4}\,\frac{1}{|w_0|+\frac{1}{2}w}+\frac{1}{4}\,\frac{1}{|w_1|+\frac{1}{2}w}\right] \qquad \text{.....(5).}$$

Here M is the mass of either particle, w the kinetic energy of the neutron, w_0 the energy of the ^{3}S state of the compound nucleus which is formed (permanently, in the case of absorption, and temporarily in the case of scattering) if the collision occurs with spins parallel, and w_1 the energy of the ^{1}S state in which the spins are antiparallel. It will be noted that only the numerical values and not the signs of the energy levels are involved. Now the deuteron is formed by the union of proton and neutron with parallel spins, and its binding energy is $2{\cdot}2\times10^6$ ev.; thus

$w_0 = -2{\cdot}2 \times 10^6$ ev. If w is small compared with w_0 and w_1, the expression for the scattering cross-section reduces to

$$\sigma_{sc} = \frac{h^2}{\pi M}\left[\frac{3}{4\,|\,w_0\,|} + \frac{1}{4\,|\,w_1\,|}\right] \qquad \text{.....(6),}$$

which is independent of w. Thus unless w_1 is very near zero, the scattering cross-section of free protons for slow neutrons should be independent of velocity.

For neutrons of thermal energies, however, the binding of the protons must be taken into account, which may be done in outline by treating the bound proton as an oscillator of frequency ν and considering the various possible types of collision, namely, (1) elastic collisions, (2) collisions in which the neutron loses energy $h\nu$, (3) collisions in which the neutron loses energy $2h\nu$, and so on. The partial cross-section for elastic collisions is found to fall monotonically (curve 0 in figure 1).

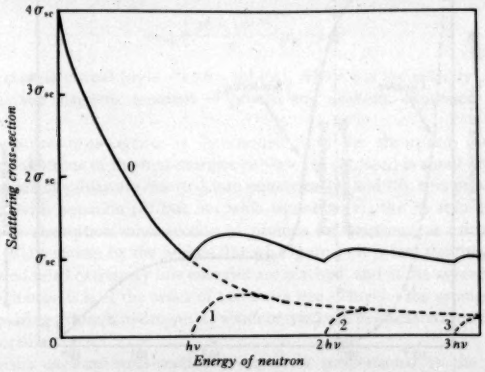


Figure 1. Cross-sections for scattering of neutrons by proton oscillators of frequency γ.

The cross-section for energy loss $h\nu$, which is, of course, zero for energies below $h\nu$, rises to a maximum and then falls away towards zero (curve 1); similarly for the other partial cross-sections. The total cross-section is shown by the full line and at high energies (much greater than $h\nu$) approaches a constant value of one-quarter the value for zero velocity. This constant value must be σ_{sc} as given by equation (6). There will actually be more than one natural frequency, but for each there will be this four-fold change in the scattering cross-section. The free path for scattering of neutrons in paraffin should therefore rise by a factor of nearly four as we go from thermal velocities to energies of a volt or so, and should thereafter remain constant until the energy of the neutron approaches the binding energy of the deuteron or of the anti-parallel neutron-proton combination, whichever is numerically the smaller.

The free path λ for scattering of neutrons may be measured directly by interposing layers of paraffin between a source of slow neutrons and a detector, or

indirectly by several methods, of which one of the most elegant was devised by Amaldi and Fermi[18] and consists in comparing the concentrations of neutrons at various distances inside the plane surface of a large paraffin block through which neutrons are diffusing. The theory of the diffusion process shows that the concentration at the surface will be finite and will increase linearly with depth below the surface in such a way that the extrapolated graph of concentration against position crosses the zero at a distance $\frac{2}{3}\lambda$ outside the surface. We shall see later that by using suitable detectors in combination with filters of cadmium it is possible to make measurements upon certain fairly well defined portions of the velocity-spectrum; for those (group C) which are mainly of thermal energy this method gave a free path of 0·3 cm., in agreement with the direct method. For neutrons

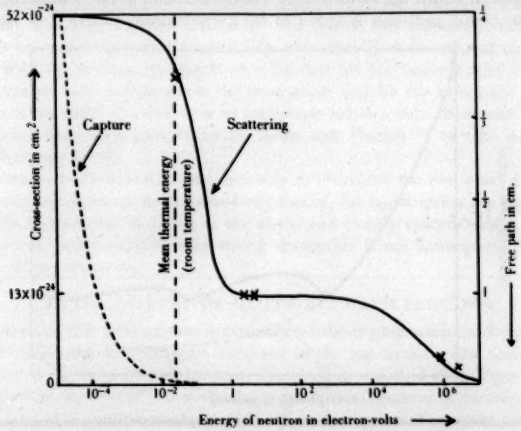


Figure 2. Capture and scattering cross-sections of protons in paraffin wax.

(group D) whose energy is believed to be about 2·4 ev., the free path came out as 1·6 cm. Another method gave $\lambda = 1{\cdot}1$ cm. for group D and the same value for another part of the velocity spectrum, group A, whose energy is about 6 ev. These figures substantiate, as well as could be hoped, the theoretical prediction that the free path for neutrons of a few volts' energy should be constant at a value four times that for very slow neutrons; the value $\lambda = 1{\cdot}1$ cm. corresponds to a scattering cross-section of 13×10^{-24} cm.^{2} Equation (6) then gives $|w_1| = 1{\cdot}16 \times 10^6$ ev. Putting this value in equation (5) and making allowance for the binding of the proton as described above, we may now draw the complete curve showing scattering cross-section as a function of the kinetic energy of the neutron. Such a curve is shown in Figure 2, and on it are marked experimental points obtained by various observers. The agreement between theory and experiment is satisfactory.

Absorption cross-section. The collision radius for the capture of a neutron by a proton is certainly no more than 10^{-12} cm., and a simple calculation shows the orbital angular momentum of the compound nucleus formed by the union of the proton with a slow neutron to be small compared with $h/2\pi$. The neutron, if it is captured, must therefore go first into an S state. The stable state of the deuteron is also an S state, and therefore the γ ray emitted on the transition to the ground state cannot be the ordinary electric dipole radiation which carries away angular momentum but will be of the magnetic dipole type. On this assumption Fermi calculated the capture cross-section; his result may be written in the form

$$\sigma_{\text{cap}} = \frac{7{\cdot}8 \times 10^{-22}}{v} (\mu_p - \mu_n)^2 \qquad \text{.....(7),}$$

if the ^{1}S state of the neutron-proton combination is stable ($w_1 = -1{\cdot}16 \times 10^6$ ev.) and

$$\sigma_{\text{cap}} = \frac{20 \times 10^{-22}}{v} (\mu_p - \mu_n)^2 \qquad \text{.....(8),}$$

if the ^{1}S state is virtual ($w_1 = +1{\cdot}16 \times 10^6$ ev.). Here v is the velocity of collision, μ_p and μ_n the magnetic moment of proton and neutron, expressed in nuclear magnetons.

The capture cross-section as determined from the absorption coefficient of paraffin for neutrons of thermal energies ($v \sim 2 \times 10^5$ cm./sec.) is about 4×10^{-25} cm.^{2} When numerical values are inserted into equations (7) and (8), this proves to be in agreement with equation (8) but not with equation (7); the ^{1}S state is therefore virtual. The absorption cross-section of protons for neutrons, as calculated from equation (8), is shown by the broken line of Figure 2; it is less than the scattering cross-section until extremely low energies are reached, and at the average energy of thermal agitation it is of the order of 100 times less. Thus on the average a thermal neutron passing through hydrogen (or water or wax) will be scattered 100 times before it is captured.

Since the capture cross-section is inversely proportional to the velocity of collision, and the number of protons met per second is directly proportional to the relative velocity, it follows that a neutron has a constant mean lifetime in hydrogen depending only upon the number of protons per unit volume. For paraffin or water this comes out to be about 2×10^{-4} sec.

The velocity-distribution of neutrons in paraffin wax. When fast neutrons are generated within a block of wax, an equilibrium will be set up in which the rate of generation of fast neutrons is balanced by the capture of slow neutrons by the protons. At energies above about a volt, the protons may be treated as free and absorption may be entirely neglected; it is then easy to calculate the equilibrium distribution of velocities. If N fast neutrons are generated per second with velocity v, the average number of neutrons existing with velocities between v and $v + dv$ is found to be

$$N(v)\,dv = \frac{2S\lambda}{v^2}\,dv \qquad \text{.....(9),}$$

where λ is the free path at the velocity v, and will be constant over a wide range of velocities. In the thermal region of velocities the distribution will be very nearly Maxwellian, and the total number will be equal to the rate of generation S multiplied by the mean life τ:

$$N_{\mathrm{Th}} = S\tau \qquad \text{.....(10).}$$

In the intermediate region accurate calculations are difficult, but it is probable that the increased scattering cross-section due to the binding of the proton will be outweighed by the smaller exchange of energy at each collision due to the increased effective mass of a bound proton, and that the neutrons will spend more time in this region of energy than if the protons were free. A full discussion is given by Bethe[19].

The flux of neutrons across unit area, which is often more important than the actual number of neutrons present, may be calculated simply if we assume the fast neutrons to be distributed uniformly through an infinite block of wax. If S now be the rate of generation of neutrons *per cm.*^{3}, the number of neutrons with velocities between v and $v+dv$ crossing unit area (both sides) per second is

$$M(v)\,dv = \tfrac{1}{2}v.\ \ N(v)\,dv = \frac{S\lambda}{v}\,dv \qquad \text{.....(11),}$$

and the flux over a finite velocity range, v_1 to v_2, is

$$\int_{v_1}^{v_2} M(v)\,dv = S\lambda \log_e \frac{v_2}{v_1}$$

$$= \tfrac{1}{2}S\lambda \log_e \frac{E_2}{E_1} \qquad \text{.....(12),}$$

where E_1, E_2 are the kinetic energies of neutrons of velocities v_1 and v_2. Thus for energies above about a volt the flux is proportional to the logarithmic width of the range of energies included.

§4. RESONANCE ABSORPTION BANDS

The most notable experimental advances made during the last year have been founded on the discovery, made independently by Amaldi and Fermi[18] in Rome and by Szilard[20] in Oxford, that the selectivity in the absorption of slow neutrons becomes much more pronounced if the neutrons are filtered through a layer of cadmium. Since this is true for nearly all pairs of absorber and detector, it follows that a certain part of the velocity-spectrum is absorbed by all elements (and particularly strongly by cadmium), while different elements absorb different parts of the rest of the spectrum. The neutrons strongly absorbed by cadmium have been called group C. Since neutrons filtered through cadmium are practically unaffected by changes of temperature, the absorption band of cadmium must cover the thermal region of velocities.

Amaldi and Fermi found remarkably high absorption coefficients of some elements for the neutrons which they absorb selectively; for instance, absorption curves for silver, made with a silver detector upon neutrons previously filtered through cadmium, could be analysed into two portions, group A having an absorption coefficient of 20 cm.^{2}/g., and group B with an absorption coefficient of

0·3 cm^2/g. Both these figures are higher than the absorption coefficient for thermal neutrons (group C).*

Fermi and Amaldi, and other workers, found similar selective absorption bands in a number of elements; occasionally the bands overlap, as, for instance, group D, which is strongly absorbed in rhodium, iridium, and indium, but not, so far as is known, in other elements. A list of some known absorption bands is given in the accompanying table.

Table I. Energies of resonance absorption bands (Goldsmith and Rasetti).

Element ...	As	Br	Rh	Ag	Ag	In	I
Half-life of isotope formed by absorption of neutron	26h.	18m.	44s.	22s. (Group A)	22s. (Group B)	54m.	25m.
Energy of absorption band (electron-volts)	84*	62*	1·16* 0·98†	2·5*	4·5*	1·38*†	80

* From absorption coefficient in boron. † From absorption coefficient in lithium.

Each group will not, of course, really be homogeneous in its absorption coefficient, for that would imply a straight-sided band in the curve connecting absorption cross-section with velocity; nevertheless, this proved at first a convenient approximation to make.

The many-body theory of nuclear reactions, described above, readily interprets these bands as resonance absorptions. Owing to the $1/v$ factor the absorption cross-section will, *caeteris paribus*, be greatest when a band happens to fall in the region of very low velocities, as appears to be the case for cadmium. We may expect the spacings of the nuclear resonances to be of similar magnitude for elements of similar atomic number, but since the energy of excitation after the absorption of a neutron of zero velocity is some millions of volts, the value of the smallest kinetic energy of the neutron which will make the total energy equal to one of the nuclear levels will vary in a random way from one nucleus to another. It may be remarked that a nuclear resonance which corresponds to a small *negative* kinetic energy of the incident neutron will also produce strong absorption of very slow neutrons, for although the peak of the absorption band will be virtual the tail of the band will extend into the region of positive energy.

Energies of the bands. Of the various absorption bands so far identified, the C band is certainly that of lowest energy because it contains the thermal neutrons; since thermal neutrons are in the majority, experiments done with group C must be nearly equivalent to experiments upon thermal neutrons alone, even if the peak of the cadmium absorption is somewhat above thermal energies.

The simplest and most unimpeachable method of arranging the groups in order of energy was used by Amaldi and Fermi[18] and consists in comparing very carefully the spatial distribution of the neutrons of the various groups around a source immersed in water; the slower the neutron the further on the average it will have diffused from the source. The order of ascending energy was found to be C, D, (A+B), I.

* The term "group" merely means "absorption band"; it does not imply that the velocity spectrum of the incident neutrons contains a large number of neutrons having a particular velocity.

Substantially the same order was found by v. Halban and Preiswerk[21,22], who showed that neutrons of (e.g.) group I could be converted by scattering in paraffin into group C, but not vice versa.

According to the many-body nuclear theory, the $1/v$ law will hold very accurately for the capture of slow neutrons by boron and lithium; by comparing the absorption coefficient of either of these elements for thermal neutrons (whose average energy is known) with its absorption coefficient for the neutrons selectively absorbed by any element, the energy at which the absorption band lies may be found[23,24]. The essential correctness of this method seems certain, for (1) the order found is the same as is given by the methods described above, and (2) the actual energies found are the same whether boron or lithium is employed. Values obtained by this method by Goldsmith and Rasetti[25] are given in the table.

Widths of the bands. Amaldi and Fermi[18] have measured the widths of several resonance bands by a method depending on the following principles. Let us assume the band to be sharp-sided, extending from an energy E_1 to an energy E_2. If S fast neutrons are generated per cm.^{3}/sec. within a large volume of water the flux of "resonance neutrons" is, by equation (12), $\frac{1}{2}S\lambda \log_e E_2/E_1$ per cm.^{2}/sec. This flux can be found experimentally from the activity induced in a thin specimen of the selectively absorbing elements in question, its dimensions and its resonance absorption coefficient being known. During the measurement thermal neutrons will be screened off by cadmium. Since λ is known and S can be found (by means of equation (10) or its equivalent) by measuring the flux of thermal neutrons with a detector sensitive only to group C, the logarithmic width of the absorption band may be obtained.

In practice, measurements have to be made in the non-uniform distribution of neutrons around a source immersed in a large volume of water, and corrections must be applied for the finite thicknesses of the specimens, so that both the theoretical and the practical details are complicated. It is desirable to make the two sets of measurements with the same specimen, first bare and secondly shielded with cadmium.

If only one absorption band is effective, the absorption coefficient for thermal neutrons will arise solely from the extension of the tail of this band into the thermal region, and will depend both on the position and on the width of this band. The position being known, the width may be calculated from the ratio of the resonance cross-section to the thermal cross-section, by means of the one-level formula (equation (1)).

Values of the widths obtained in these ways range from about a tenth of a volt to 2 v.; for details and for a discussion of the reliability of the results, reference should be made to pp. 141–5 of Bethe's Report[13].

The measured width is nearly always considerably larger than the Doppler width due to thermal agitation of the absorbing nuclei. Comparison of the resonance cross-sections with equation (3) shows, in agreement with the small amount of scattering which is observed, that Γ_n/Γ is small; the measured width is therefore principally a γ-ray width.

For the A group of Ag 22″, however, the observed width is almost exactly equal to the Doppler width, which means that the natural width is small compared with the Doppler width. Bethe concludes that the natural width is less than 0·03 v.; since this must also be the width of the nuclear level into which the neutron is captured, and since the energy of this level above the ground state is probably more than a million volts, it must be sharp to about one part in 10^8.

§5. SPINNING DISCS AND MECHANICAL VELOCITY SELECTORS.

The velocities of thermal neutrons ($\sim 2 \times 10^5$ cm./sec.) are not many times greater than speeds attainable mechanically, and a number of experiments depend upon the use of high-speed moving parts. The first successful experiments of this kind were made in Rome[26], but the most important subsequent work has been done in New York, where two types of experiment have been performed.

If the $1/v$ absorption law is true, the mean life of a neutron in any medium is independent of any relative motion of the neutron and the absorbing nuclei, because any increase in the number of nuclei encountered per second is exactly offset by the decreased chance of capture at each encounter. If therefore a beam of neutrons falls obliquely upon a spinning disc of absorbing material, the number of neutrons transmitted will be independent of the motion if the $1/v$ law is true, while if the variation of cross-section with velocity were less rapid (e.g. independent of v) fewer would be transmitted when the disc spins towards the oncoming neutrons (because each neutron meets more atoms) and more would be transmitted when the disc is spinning away from the neutrons. Using a peripheral speed of $\pm 1{\cdot}4 \times 10^4$ cm./sec., Rasetti, Segrè, Fink, Dunning and Pegram[27] found no effect with a steel disc coated with silver, but a disc coated with cadmium transmitted 8 per cent fewer slow neutrons when moving against than with the beam. This is about the change which would be expected if the cross-section of cadmium were independent of velocity. A later experiment done with boron showed no change, and the $1/v$ law is therefore confirmed (at least for the slower neutrons, for which the method is clearly most sensitive) for silver and boron. The result found for cadmium must mean that the resonance band lies within the thermal region, which is in accordance with the small effect of temperature upon neutrons filtered through cadmium. Segrè and Zahn conclude that the data can be explained by assuming the peak of the absorption band to lie at $0{\cdot}05_8$ v., with a width of about 0·05 v.

By means of a duralumin disc coated with sectors of cadmium and rotating at high speed in front of a similar stationary disc, Dunning, Pegram, Fink, Mitchell and Segrè[28,29] were able to interrupt a beam of slow neutrons, and to measure their times of transit to a second shutter whose moving element was on the same shaft as the first, so that the two shutters opened and closed with the same frequency but with a phase relation which could be varied by changing the orientation of one of the fixed discs. Owing to the imperfect opacity of cadmium, the clear portions of the shutters cannot be made a very small fraction of the whole, or the neutrons passing through the openings would be swamped by the background. The resolution obtainable is therefore only low, but was sufficient to demonstrate that the velocity-

distribution of the neutrons leaving a paraffin howitzer is roughly Maxwellian, and that it changes in the proper direction when the source is cooled to liquid air temperature. An improved machine has been built and is described by Fink[30].

§6. THE DIFFRACTION OF NEUTRONS

The de Broglie wave-length of a neutron of average thermal velocity is about $0{\cdot}6 \times 10^{-8}$ cm., which is of the order of the lattice constants of crystals. Diffraction effects may therefore be expected when neutrons are elastically scattered from crystalline bodies. The clearest evidence for diffraction of neutrons which has so far been obtained is due to Mitchell and Powers[31], who used the arrangement shown diagrammatically in Figure 3. Neutrons from a paraffin howitzer, which gives a certain amount of collimation of the neutron beam and considerably increases the intensity obtainable, were scattered from single crystals of MgO arranged around a circle, and detected by a boron chamber on the axis of the system. The geometry of the apparatus and the orientation of the crystals were such that the peak of the Maxwell distribution would be diffracted at the Bragg angle from the (100) planes of the crystals to the detector. Absorbers and wax were arranged so as to prevent, as far as possible, unwanted neutrons from reaching the detector. When the crystals were turned from the Bragg position, alternate crystals being turned through $\pm 45°$ so that the average distribution of scattering material was very little altered, a small but definite decrease was observed in the number of neutrons recorded. A check on the reality of the effect was made by replacing the single crystals by polycrystalline scatterers of aluminium, and observing that there was no such change when these were rotated.

§7. THE MAGNETIC MOMENT OF THE FREE NEUTRON

The magnetic moments of the proton and the deuteron have been measured experimentally and are about $+2{\cdot}7$ and $+0{\cdot}8$ nuclear magnetons respectively. The positive sign indicates that the directions of the magnetic and mechanical moments are related in the same way as for a rotating positive charge. If nuclear magnetic moments are additive, then since the deuteron is formed by the union of proton and neutron with their mechanical moments parallel, the magnetic moment of the neutron will be $(0{\cdot}8-2{\cdot}7)$ or about -2.

Bloch[32] showed theoretically that the scattering of neutrons in magnetized iron will depend perceptibly upon the orientation of their magnetic moments with respect to the direction of magnetization, so that a beam of neutrons will be partially polarized after passage through a plate of iron magnetized at right angles to the direction of the beam, and the number of neutrons transmitted through two plates in succession will be greatest when the directions of magnetization are parallel. Effects of this kind, of the order of magnitude predicted, have been observed by various workers. Since the effects are small (2 or 3 per cent) and the details of the theory are not yet very certain, such experiments can be considered only as establishing the existence of the phenomenon; they cannot in any case determine the sign of the magnetic moment.

If in the gap between the two magnetized plates there exists a magnetic field *parallel* to the neutron beam, the neutrons will precess and the plane of polarization

of the beam on reaching the analyser plate will have rotated in a direction and by an amount which depend upon the sign and magnitude of the magnetic moment. Experiments depending on this principle have been made by Frisch, v. Halban and Koch[34], and have shown the moment to be negative, and of the order of -2, as is required by the assumption that the magnetic moment of the deuteron is derived by the (vector) addition of the moments of proton and neutron. This method has the advantage of being independent of ferromagnetic theory, since the magnetized plates are used only as instruments. The longitudinal field which causes precession is only a few gauss and is produced by a solenoid in air. A similar result has been found by Powers, Carroll Beyer and Dunning[33] by another precession method.

REFERENCES

(1) Fermi, E., Amaldi, E., D'Agostino, O., Rasetti, F. and Segrè, E. *Proc. roy. Soc.* A, **146**, 483 (1934).
(2) Amaldi, E., D'Agostino, O., Fermi, E., Pontecorvo, B., Rasetti, F. and Segrè, E. *Proc. roy. Soc.* A, **149**, 522 (1935).
(3) Moon, P. B. and Tillman, J. R. *Proc. roy. Soc.* A, **153**, 476 (1936).
(4) Chadwick, J. and Goldhaber, M. *Nature, Lond.*, **135**, 65 (1935).
(5) Amaldi, E. *Nuovo Cim.* **12**, 223 (1935).
(6) Bonner, T. W. and Brubuker, W. M., *Phys. Rev.*, **49**, 778 (1935).
(7) Burcham, W. E., and Goldhaber, M. *Proc. Camb. phil. Soc.* **32**, 632 (1936).
(8) Dunning, J. R., Pegram, G. B., Fink, G. A. and Mitchell, D. P. *Phys. Rev.* **48**, 265 (1935)
(9) Bjerge, T. and Westcott, C. H. *Proc. roy. Soc.* A, **150**, 709 (1935).
(10) Bethe, H. A. *Phys. Rev.* **47**, 747 (1935).
(11) Bohr, N. *Nature, Lond.*, **137**, 344 (1936).
(12) Bohr, N. and Kalckar, G. *Det Kongelige Danske Videnskabernes Selskabs, Mathematisk-fysiske Meddelelser.* (In Press.)
(13) See also Bethe, H. A. *Rev. Mod. Phys.* **9**, No. 2 (April 1937).
(14) Breit, G. and Wigner, E. *Phys. Rev.* **49**, 519 (1936).
(15) Bethe, H. A. and Placzek, G. *Phys. Rev.* **5**, 450 (1937).
(16) Peierls, R. *Rep. Prog. Phys.* **2**, 27 (1936).
(17) Fermi, E. *Ric. Scient.* Anno VII, **2**, 13 (1936).
(18) Amaldi, E. and Fermi, E. *Phys. Rev.* **50**, 899 (1936).
(19) Bethe, H. A. *Rev. mod. Phys.* **9**, No. 2, 126 (April 1937).
(20) Szilard, L. *Nature, Lond.*, **136**, 950 (1936).
(21) Von Halban, H. and Preiswerk, T. *Nature, Lond.*, **137**, 905 (1936).
(22) Von Halban, H. and Preiswerk, P. *Helv. phys. Acta*, **9**, 318 (1936).
(23) Frisch, R. O. and Placzek, G. *Nature, Lond.*, **137**, 357 (1936).
(24) Weekes, D. F., Livingston, M. S. and Bethe, H. A. *Phys. Rev.* **49**, 471 (1936).
(25) Goldsmith, H. H. and Rasetti, F. *Phys. Rev.* **50**, 328 (1936).
(26) Amaldi, E., D'Agostino, O., Fermi, E., Pontecorvo, B. and Segrè, E. *Ric. Sci.* Anno VI, **1**, 583 (1935).
(27) Rasetti, F., Segrè, E., Fink, G. A., Dunning, J. R. and Pegram, G. B. *Phys. Rev.* **49**, 104 (1936).
(28) Dunning, J. R., Pegram, G. B., Fink, G. A., Mitchell, D. P. and Segrè, E. *Phys. Rev.* **48**, 704 (1935).
(29) Fink, G. A., Dunning, J. R., Pegram, G. B. and Mitchell, D. P. *Phys. Rev.* **49**, 103 (1936).
(30) Fink, G. A. *Phys. Rev.* **50**, 738 (1936).
(31) Mitchell, D. P. and Powers, P. N. *Phys. Rev.* **50**, 486 (1936).
(32) Bloch, F. *Phys. Rev.* **50**, 259 (1936).
(33) Powers, P. N., Beyer, H. G., Carroll, H. and Dunning, J. R. *Phys. Rev.* **52**, 38 (1937).
(34) Frisch, R. O., v. Halban, H. and Koch, J. *Nature, Lond.*, **139**, 756 (1937).

THE CHARGE OF THE ELECTRON*

By PROF. H. R. ROBINSON, D.Sc., Ph.D., F.Inst.P., F.R.S.
Queen Mary College (University of London)

§1. INTRODUCTION

In an article published under the same title in the *Progress Reports* for 1935[1], the main point under discussion was the difference between the X-ray value of e (then taken to be $4{\cdot}805 \times 10^{-10}$) and the older-established oil-drop value ($4{\cdot}769 \times 10^{-10}$). The ratio of these two values is almost exactly 137/136, and each value had been claimed to be correct to within 1 part in 1000. At the time when the article was written, just over 2 years ago, there was strong, but not quite irrefragable, evidence in favour of the X-ray value. Further, it seemed that the chief cause of the discrepancy (assuming it to be a real discrepancy) most probably lay in the viscosity term used in calculating the oil-drop value—that is, in errors either in the coefficient of viscosity assumed for air, or in the form of the correction applied to Stokes's law, or in both. This view, first brought into prominence by Shiba, has received experimental support during the past 2 years; at the same time, additional evidence has become available in support of the X-ray method, and it seems quite certain that the 1917 value of e must be abandoned.

§2. RECENT WORK RELATING TO THE OIL-DROP VALUE

In 1935 Kellström published a short note[2] on a careful re-determination of the coefficient of viscosity of air, reporting a value much higher than that used by Millikan. Very shortly afterwards, Bond[3] published an account of an independent measurement of this coefficient, in almost exact agreement with that of Kellström. The two values (both reduced to 23° C.) are:

$$\text{Kellström: } \eta = (1834{\cdot}8 \pm 3{\cdot}0) \times 10^{-7} \text{ poise.}$$

$$\text{Bond: } \eta = (1834{\cdot}7 \pm 0{\cdot}8) \times 10^{-7} \text{ poise.}$$

A little later, Kellström[4] published a revised value $(1834{\cdot}9 \pm 2{\cdot}7) \times 10^{-7}$.

These are just over 0·65 per cent in excess of Millikan's value. As η appears raised to the (3/2) power in the expression for e, the adoption of the new value raises the oil-drop e by almost exactly 1 per cent—that is, to $4{\cdot}817 \times 10^{-10}$, which is even larger than the X-ray value.

Meanwhile, however, Bäcklin and Flemberg had undertaken a complete repetition of the oil-drop experiments, using a wide range of gas pressures in order to

* Throughout this *Report*, when e appears alone, it must be understood to be measured in absolute e.s.u. When it appears in the ratio e/m_0, it is measured in absolute e.m.u. To save repetition, these units are generally omitted in the text. The practice of omitting the factors 10^{-10} and 10^3 has however not been followed.

test more thoroughly the deviations from Stokes's law. According to their preliminary statement[5], there are errors in Millikan's measurements; the new oil-drop results, combined with the new values of η, give $e = 4{\cdot}800 \times 10^{-10}$. The nine separate determinations tabulated by Bäcklin and Flemberg are very concordant (average difference from the mean value $\pm 0{\cdot}008 \times 10^{-10}$), and apart from possible errors in η the final result should be accurate to well within 1 part in 1000. The agreement with the latest X-ray value is therefore satisfactory.*

§3. RECENT WORK ON THE X-RAY VALUE

(*a*) *General.* Confidence in the X-ray value of e is increased by the work of DuMond and Bollman[6]. In this, the diffraction angles were measured for X-rays diffracted from finely powdered calcite; the density of the calcite powder was measured accurately by standard methods, and an X-ray value of e was deduced from the measurements. The result—$(4{\cdot}799 \pm 0{\cdot}007) \times 10^{-10}$—is in excellent agreement with the other X-ray values. The important point in this experiment is that the grains of calcite powder had linear dimensions comparable with the thickness of the surface layer which is effective in the ordinary (Bragg) type of reflection. This disposes effectively of a serious objection to the X-ray e, viz. that the Bragg reflection is almost a pure surface effect, while the density is measured for a large crystal. The arguments which could be opposed to this objection in the 1935 *Report* (pp. 252–3) were strong, but perhaps not quite so conclusive as this result of DuMond and Bollman—certainly nothing like so direct.

(*b*) *Numerical.* In the 1935 *Report* the following were quoted as the most reliable values of e (X-ray):

Bäcklin[7]: $4{\cdot}805 \times 10^{-10}$ ($\pm 0{\cdot}04$ per cent).
Södermann[8]: $4{\cdot}806 \times 10^{-10}$ ($\pm 0{\cdot}03$ per cent).
Bearden[9]: $4{\cdot}8036 \times 10^{-10}$ ($\pm 0{\cdot}01$ per cent).

Readers who consult the original papers should take warning that none of these authors employs the conventional Moseley-Siegbahn wave-length scale proposed by Siegbahn, in spite of the very real advantages in its use. Bäcklin and Södermann use a scale differing from it by about 1 part in 7400, while Bearden's scale differs from it by about 1 in 2250 in the same sense. The differences are not perhaps enormous, but they are big enough to be disconcerting, if unremarked, to anyone working over the results with any aid to calculation much more powerful than, for instance, three-figure logarithms.

There is one slight correction now to be applied to the results of Bäcklin and Södermann. Their values of e were based on comparisons of the ruled grating and crystal wave-lengths of the aluminium $K\alpha$ doublet, and for the crystal wave-length they took Larsson's[10] value, published in 1929. It seems certain from the later work of Haglund[11] that Larsson's value was too low by about 1 part in 5000. Bäcklin's and Södermann's values of e should therefore be reduced by approximately 1 part in 1700.

* But see note at end, added in proof.

This brings the mean of their values to about 1 part in 5500 *below* the result given by Bearden. Bearden's value, however, is a weighted mean of his own, Bäcklin's and Södermann's individual determinations, in which the first predominate. Bearden's e therefore must also be very slightly reduced, and we may take in round figures:

$$e\ (\text{X-ray}) = 4{\cdot}803 \times 10^{-10}.$$

The agreement between the three sets of measurements, when the new value of the crystal wave-length of aluminium $K\alpha$ is used, is so convincing that we are probably justified in regarding this* as correct to within 0·01 or 0·02 per cent.

§4. ELECTRON-DIFFRACTION VALUE OF e

Recently von Friesen[12], as the result of a very thorough and carefully planned investigation of the diffraction of electrons at the surface of galena crystals, has arrived at a value of $e = (4{\cdot}796 \pm 0{\cdot}005) \times 10^{-10}$.

In calculating this, von Friesen assumed for e/m_0 the value $1{\cdot}7592 \times 10^7$, which may be rather high (cf. § 6, *infra*). If we take $e/m_0 = 1{\cdot}758 \times 10^7$, von Friesen's e becomes $(4{\cdot}797 \pm 0{\cdot}005) \times 10^{-10}$.

The agreement with the X-ray and new oil-drop values is fair, and it appears that von Friesen has pushed the accuracy of electron diffraction measurements very nearly to the limit. The method however does not yield a very direct measurement of e; apart from the necessity of assuming a value for e/m_0, there are two major assumptions—namely, that de Broglie's formula for electron wave-lengths and Bohr's formula for the Rydberg constant are exact relations. (The Rydberg constant is introduced in order to eliminate h from the equations, in favour of constants which are more accurately known.)

§5. THE BEST VALUE OF e

To sum up the changes since the 1935 *Report*—the X-ray values of e obtained by different workers are seen to be in even better agreement than appeared at first, and one grave objection to the X-ray method has been satisfactorily dismissed. Unless, therefore, there is something wrong with the theory (or with the accepted value of the Faraday), the X-ray value† $4{\cdot}803 \times 10^{-10}$ is by far the most reliable that we have.

The oil-drop work is subject to more, and more serious, sources of experimental error than the X-ray work, and it seems unlikely that the accuracy of the oil-drop measurements can be much further improved. The new measurements are, however, as has already been mentioned, in substantial agreement with the X-ray results.

Birge, whose penetrating analyses of a vast range of experimental data have made him a sort of supreme arbiter on these problems, now adopts[13, 14]

$$e = (4{\cdot}8029 \pm 0{\cdot}0005) \times 10^{-10}$$

as the best value.

* Regarded purely as the X-ray value, of course.

† $4{\cdot}805 \times 10^{-10}$ was the value given in the 1935 *Report*.

§6. GENERAL

The present agreement between the oil-drop and the X-ray results is in itself very satisfactory. If, however, we examine the results of certain other experiments, we find less cause for complacency. This matter, to which we shall return shortly, has been well discussed by Birge in the references already given[13, 14], and elsewhere.

Further, by another curious shift (compare the 1935 *Report*, p. 247), our knowledge of the value of e/m_0 seems now to be less precise than it was thought to be two years ago. Then, $1{\cdot}758 \times 10^7$ seemed to be a very well-established value. The best measurements between 1932 and 1935 ranged only from $1{\cdot}7570 \times 10^7$ (one "spectroscopic"* and one "deflection" value) to $1{\cdot}7587 \times 10^7$ (deflection). The value adopted by Birge in 1936[13], $(1{\cdot}75762 \pm 0{\cdot}00026) \times 10^7$, was the mean of five recent measurements, made by four distinct methods.

Now, however, Dunnington[15] reports the "deflection" value

$$(1{\cdot}7597 \pm 0{\cdot}0004) \times 10^7$$

as the result of a very elaborate investigation, and Houston[16] gives

$$(1{\cdot}7601 \pm 0{\cdot}0015) \times 10^7,$$

obtained by applying more powerful methods of analysis to the fine structure of the Balmer lines of hydrogen and deuterium. (The latter result is not given in this form in Houston's paper, but $1{\cdot}7601 \times 10^7$ is the mean of the two results which he gives, and the probable error is that quoted by Dunnington[15] on Houston's privately communicated authority.)

On the other hand, Shaw[17] has made a preliminary report of new deflection measurements which give $e/m_0 = (1{\cdot}7571 \pm 0{\cdot}0013) \times 10^7$. The combined effect of the new measurements is therefore to put us back pretty well to where we were two years ago, but with an enhanced sense of insecurity. This is of some importance in what follows.

To return now a little closer to the region prescribed by the terms of reference of this *Report*: the major difficulty raised by the adoption of the new value of e appears in this way:

If Bohr's formula for the Rydberg constant is exact, we may write:

$$\frac{h}{e} = \sqrt[3]{\frac{2\pi^2 e^3}{Rc^2 e/m_0}}.$$

It is customary to assume that errors in R and c are negligible compared with those in e and e/m_0.

If now (cf. Birge[13, 14]) we take $e = 4{\cdot}803 \times 10^{-10}$, $e/m_0 = 1{\cdot}758 \times 10^7$, $R = 109737$, $c = 2{\cdot}998 \times 10^{10}$, we get

$$h/e = 1{\cdot}3797 \times 10^{-17} \text{ erg-sec.-e.s.u.}^{-1},$$

and this result is very insensitive to small changes in e/m_0. It ought therefore to be very accurate if our value of e is reliable.

* Zeeman effect.

Now the examination of the inverse photoelectric effect (measurement of the short-wave limit of the continuous X-ray spectrum excited by cathode rays which have fallen through a known potential difference) appears to be by far the most accurate direct method of measuring h/e. There have been three recent determinations by this method[18, 19, 20]. They are in good agreement with one another, and we need only quote the latest, that of DuMond and Bollman[20], published in 1937. These experiments, carried out with the most extreme care, and with elaborate precautions, gave $h/e = (1{\cdot}3762 \pm 0{\cdot}0003) \times 10^{-17}$ (assuming the ruled-grating values for the X-ray wave-lengths). This is a great deal lower than the value just deduced from Rydberg's constant, and the other two values[18, 19] are slightly lower still.

This is serious, and we might well, adapting a phrase from Kelvin, speak of a "twentieth-century cloud over our knowledge of the atomic constants". So long as this discrepancy remains unexplained, we cannot have complete confidence in the present value of e. The matter has, of course, been discussed by Birge[13, 14]; possible explanations—almost equally unattractive—are (1) the X-ray e is wrong, (2) the undoubtedly accurate and carefully considered experiments on the short-wave limit are subject to some unknown defect, or (3) Bohr's formula for R is merely an approximation.

It will no doubt be noticed that the two values for h/e which have just been quoted differ by very nearly $\frac{1}{4}$ per cent—that is, they stand very nearly in the ratio $\sqrt[4]{(137/136)}$. There is, however, no clear-cut evidence of any serious difference between the spectroscopic and deflection values of e/m_0.

There remains to be considered the work on the energies of the photoelectrons expelled from metals by monochromatic X-rays (cf. 1935 *Report*, § 5(*c*)). This has recently been analysed in some detail[21], and it yields the result

$$h/e = e/m_0 \times 7{\cdot}844 \times 10^{-25}.$$

Taking the fairly unobjectionable value $1{\cdot}758 \times 10^{7}$ for e/m_0, we get

$$h/e = 1{\cdot}379 \times 10^{-17} \text{ erg-sec.-e.s.u.}^{-1}$$

—which is in very close agreement with the value deduced from Rydberg's constant. Here, of course, we get the full proportional effect of any error in the value assumed for e/m_0. The value which would bring these results into exact agreement with the Rydberg constant is $1{\cdot}7589 \times 10^{7}$, which is quite a possible value.

The X-ray photoelectric results do not however give a very direct measurement of h/e, and nothing approaching the accuracy of X-ray spectroscopic measurements can be claimed for them. The cumulative error in the final results quoted may be quite large, and the agreement fortuitous; in fact, as one of those mainly responsible for the X-ray photoelectric measurements, I am quite simply unable to decide whether this agreement is a matter for self-congratulation, or a clear proof of their lack of precision.

§ 7. CONCLUSION

There is a greater weight of evidence in favour of the X-ray value of e than there was two years ago. There is much reason for thinking that the value

$$e = 4{\cdot}803 \times 10^{-10} \text{ e.s.u.}$$

may be correct to within 1 part in 10,000, and there is no longer much serious talk of two values of e differing in the ratio 136/137. There are, however, certain outstanding discrepancies among the values of related constants, and until these are cleared up we cannot accept any value of e with complete confidence.

Note added in proof, November 1937: Houston has just published the results of a new determination [22], by the rotating cylinder method, of the viscosity of air. His value for η at 23° C. is $(1829{\cdot}2 \pm 4{\cdot}5) \times 10^{-7}$ poise, which, with the stated probable errors, is not inconsistent with Kellström's [2], [4]. If we take Houston's mean value for η, Millikan's oil-drop e becomes $4{\cdot}796 \times 10^{-10}$, in fair agreement with the X-ray value. Bäcklin and Flemberg's result [5], however, comes down to $4{\cdot}778 \times 10^{-10}$—or very nearly to Millikan's 1917 value! This is a most unfortunate note on which to end the *Report*.

REFERENCES

(1) *Reports on Progress in Physics*, **2**, 247 (1935).
(2) Kellström. *Nature, Lond.*, **136**, 682 (1935).
(3) Bond. *Nature, Lond.*, **137**, 1031 (1936); *Proc. phys. Soc.* **49**, 205 (1937).
(4) Kellström. *Phys. Rev.* **50**, 190 (1936).
(5) Bäcklin and Flemberg. *Nature, Lond.*, **137**, 655 (1936).
(6) DuMond and Bollman. *Phys. Rev.* **50**, 524 (1936).
(7) Bäcklin. *Nature, Lond.*, **135**, 32 (1935); *Z. Phys.* **93**, 450 (1935).
(8) Södermann. *Nature, Lond.*, **135**, 67 (1935); *Nova Acta Soc. Sci. upsal.* (Ser. 4), IX, No. 8 (1935).
(9) Bearden. *Phys. Rev.* **48**, 385 (1935).
(10) Larsson. Diss. Uppsala (1929).
(11) Haglund. *Z. Phys.* **94**, 369 (1935).
(12) von Friesen. *Uppsala Univ. Årsskr.* 1935:14.
(13) Birge. *Nature, Lond.*, **137**, 187 (1936).
(14) Birge. *Phys. Rev.* **52**, 241 (1937). This contains a criticism of a set of constants proposed by von Friesen in *Proc. roy. Soc.* A, **160**, 424 (1937).
(15) Dunnington. *Phys. Rev.* **52**, 475 (1937).
(16) Houston. *Phys. Rev.* **51**, 446 (1937).
(17) Shaw. *Phys. Rev.* **51**, 887 (1937).
(18) Kirkpatrick and Ross. *Phys. Rev.* **45**, 454 (1934).
(19) Schaitberger. *Ann. Phys., Lpz.*, **24**, 84 (1935).
(20) DuMond and Bollman. *Phys. Rev.* **51**, 400 (1937).
(21) Robinson. *Phil. Mag.* **22**, 1129 (1936).
(22) Houston. *Phys. Rev.* **52**, 751 (1937).

EXPERIMENTAL ELECTRICITY AND MAGNETISM

Edited by L. HARTSHORN, D.Sc., A.R.C.S., D.I.C.

§ 1. THE BREAKDOWN OF INSULATING MATERIALS. ELECTRIC STRENGTH

By L. HARTSHORN, D.Sc., A.R.C.S., D.I.C.

LAST year's *Progress Report* contains a survey of recent investigations into those properties of dielectrics which are recorded in terms of dielectric constant and conductivity or power factor, and it is noted that such properties are largely dependent on molecular structure and intermolecular forces. A considerable amount of work has also been carried out in recent years on the failure of electrical insulating materials under high voltage. This is a matter which closely concerns all experimentalists who have occasion to use high voltages, and although much of the work has been carried out on account of the practical importance of the subject to the electrical industry, it is not without its theoretical interest.

The tendency of any insulating material to break down is usually represented by its electric strength, which is roughly defined as that value of the electric intensity which is just sufficient to cause the material to break down or lose its insulating properties. The underlying conception is analogous to that of the fracture of materials under mechanical stress. Indeed in this connexion the electric intensity at any point in the material is often referred to as the electric stress. It is regarded as producing an electric displacement proportional to the stress and the dielectric constant of the material as long as the stress is less than a certain critical value, but when this value is passed the electric displacement increases very rapidly, the whole structure of the material becomes unstable and collapses.

If this simple notion were adequate to account for the observed facts, a few straightforward measurements of the voltage required to break down the ordinary materials used for insulating purposes would provide the technician with all the data needed for the design of high-voltage apparatus, and would at the same time give us a direct measure of the forces holding together the structure of positive and negative charges constituting materials in question. However, no such happy result is obtained. The experimental values of breakdown voltage obtained under various conditions do not yield for any given material a definite electric strength. The electric intensity required to produce breakdown in any material is apt to vary considerably from one sample to another and also to vary with such factors as the duration and rate of application of the voltage, the thickness of the material, the frequency (if the voltage is alternating), the temperature, the size and shape of the electrodes, and the nature of the surrounding medium (if the material under test is a solid). The object of much of the experimental research on the subject is to discover the causes of these variations, and to arrive finally at a value which is independent of such factors and which does represent the electric strength of the material.

Local faults. One very obvious complication in the problem of the breakdown of solid insulating material is the possibility of the existence of weak spots in the materials. After breakdown has occurred, the material is usually found to be perforated at one or more places between the electrodes, and the punctures so produced are often so small as to be scarcely visible. Wagner[1] has ascribed such breakdowns to the existence in the material of fine conducting paths or filaments, which are heated by the passage of current through them. In the stable condition all the heat generated in these paths is dissipated into the surrounding material or into the electrodes, but since the conductivity, which is considered to be electrolytic in character, increases rapidly with rise of temperature, as the voltage is steadily increased the heat generated in the filament ultimately increases so rapidly that the amount generated cannot be dissipated. The temperature of the filament rises more and more rapidly, an unstable condition is reached, and the filament burns out like a fuse wire. Assuming that the conductance of a filament is inversely proportional to d, the thickness of the specimen, and that the heat dissipated is directly proportional to d, Wagner showed that the voltage V_m at which instability occurs, i.e. at which the voltage begins to fall as the current is increased, and which is determined by the condition $\delta V/\delta I = 0$, I being the value of the current in the filament, is directly proportional to d. In other words, the material behaves as if it had a true electric strength V_m/d independent of the thickness of the sample. Such a relation has frequently been observed for very thin materials, and it is probable that the breakdown of such materials is often of this character. Observations by Hartshorn and Ward[2] on paper afford a recent example. Wagner observed the curves connecting V and I with their point of instability by making measurements on thin samples of insulating material with electrodes of certain samples of wood which possessed considerable conductivity in the longitudinal direction, but very little in a transverse direction. These electrodes served as current-limiting resistances, which protected the samples from damage when the point of instability was reached. Somewhat similar investigations have been described by Hayden and Steinmetz[3] and by Schumann[4].

Thermal instability. All known insulating materials possess a finite electrical conductivity, and therefore dissipate a certain amount of energy as heat when placed in an electric field, and the heat generated in this way may increase sufficiently to destroy the material even in the absence of weak spots. Breakdown of this character is very common with alternating voltage of high frequencies, for most insulating materials possess a power factor that varies comparatively little with frequency, so that their effective conductivity is proportional to frequency. An obvious condition for breakdown of this character is that the conductivity shall increase with the rise of temperature. This holds for ionic conductivity of the electrolytic type, and is found to hold for most materials which absorb appreciable quantities of water. It is also obvious that breakdown of this character will occur the more easily the higher the temperature, and the less readily the material is able to get rid of the heat generated by conduction to the electrodes or outer surfaces The general problem is one of considerable complexity, but since it is very con-

venient to have data which represent the tendency of material to fail owing to thermal instability of this kind, a number of simple cases have been worked out and the appropriate constants of various materials have been measured. For example, S. Whitehead[5] considers the case of a thin sheet of dielectric with plane electrodes. The equipotential and isothermal surfaces are planes parallel to the surfaces of the plate. In order that thermal equilibrium may be possible, the heat developed in any element must equal that lost by conduction, etc. We have, therefore, if v is the voltage and θ the temperature at a distance x from one electrode

$$\lambda \frac{d^2\theta}{dx^2} + \sigma \left(\frac{dv}{dx}\right)^2 = 0,$$

where σ is the electrical conductivity of the material, and λ the thermal conductivity. Also $\frac{d}{dx}\left(\sigma \frac{dv}{dx}\right) = 0$ since the current-density is constant throughout the specimen.

An analytical treatment of the general problem has been given by Fock[6,7] and applying his method to the above simple case in which, moreover, one electrode is considered to be maintained by some means at a constant temperature θ_0, while the other is thermally insulated and therefore reaches a higher temperature θ_m, Whitehead[5] obtains the expression

$$V^2 = \int_{\theta_0}^{\theta_m} (2\lambda/\sigma)\, d\theta,$$

for the applied voltage V and the temperatures when thermal equilibrium is established. Obviously, as the applied voltage V increases, the maximum temperature θ_m for which thermal equilibrium is possible, increases and may reach a critical value θ_c say, at which the material is virtually destroyed. It may melt, char, or suddenly develop considerable conductivity. Thus, as the thickness of the material is increased, the maximum permissible applied voltage does not increase indefinitely. There is in general a maximum voltage which the material can withstand, however great the thickness may be. This holds even when λ/σ is independent of temperature, but for many materials λ/σ decreases very rapidly with rise of temperature, and the maximum permissible voltage for thermal equilibrium is correspondingly reduced. For this maximum voltage, which may be regarded as a constant of any given material, Whitehead[5] gives the expression $V_m^2 = \int_{\theta_0}^{\theta_c} (8\lambda/\sigma)\, d\theta$, but since it is very difficult to define and measure θ_c, it is more convenient to write $V_m^2 = \int_{\theta_0}^{\infty} (8\lambda/\sigma)\, d\theta$. The difference in those cases in which λ/σ increases very rapidly with temperature is not important.

Using experimentally determined values of thermal and electrical conductivity, Whitehead has calculated the maximum permissible voltages as determined by the above conditions of general (as distinct from local) thermal instability for many common dielectrics. Most of the ordinary commercial organic insulating materials such as ebonite, paper, and synthetic resin boards give values of about a million volts for alternating voltage of frequency 50 c.sec. For such materials σ is usually

roughly proportional to the frequency f, so that we may write as a sufficiently good approximation $V_m \propto 1/\sqrt{f}$. It follows that, at a frequency of 50 Mc./sec., V_m is of the order of 1000 v. for these materials. The values of V_m for materials of mineral origin, such as glass, porcelain, and other ceramics, and particularly of crystals such as quartz, mica and rock salt, are very much greater than those of the organic materials. These materials have a considerable advantage in having comparatively high thermal conductivities as well as low power factors. Their very high values of V_m provide a good indication of their usefulness for work at very high frequencies.

When breakdown occurs owing to thermal instability, the conception of dielectric strength is of doubtful value. The voltage gradient, which is just sufficient to produce thermal instability when applied for a very long time, may be exceeded with safety for short periods, so that the apparent electric strength will vary with the duration of the test. It will also diminish as the thickness of the specimen is increased, and as the temperature is increased. Various observers have noted empirical laws connecting V the voltage and t the time for which it must be applied in order to cause breakdown. A recent example is that of H. E. Mason[8], who found the law $tV^N = \text{constant}$, where $N = 13$ to 14, to hold with surprising accuracy for various rubber insulated cables, the time of application t varying from about 1 to 1000 min. Perhaps the best known of such laws is that due to Peek, viz. $V = V_0\, at^{-\frac{1}{2}}$, where V_0 and a are constants. Values of electric strength quoted by engineers are usually those obtained for a time of 1 min. with samples and electrodes of a standard size and shape.

Whitehead and Nethercot[9] have described an experimental investigation of the breakdown of dielectrics in the form of thin sheets, under thermal conditions which were controlled by the use of suitable lagging. The temperature of the material was observed by means of thermocouples. Observations of effective conductance, and of temperature-rise were made over an extended period of time, the applied voltage being increased in steps, sufficient time being allowed at each step for thermal equilibrium to be established. They also made measurements of apparent dielectric strength for short-time applications of voltage. They found that organic materials, such as cellulose acetate, which normally contain moisture, broke down owing to thermal instability, particularly at high temperatures, and that the breakdown-voltage (long time) could then be estimated from the constants of the material measured in the stable condition. The fact that some of these breakdowns took the form of a sudden explosive discharge, rather than a continuous loss of insulating properties, did not invalidate the calculations. Short time tests with larger voltages gave less satisfactory agreement between theory and experiment, although the effect of temperature change could still be predicted.

The curves given by Whitehead and Nethercot to show the variation of breakdown field-strength with temperature are not continuous. There is in all cases a certain critical temperature above which the value diminishes rapidly with rise of temperature, indicating that breakdown is due to thermal instability. At lower temperatures the breakdown field-strength is comparatively independent of temperature, indicating that the breakdown is not of the thermal type. Similar results

have been obtained by other observers, and the two classes of breakdown are often referred to as *thermal* and *electric*, the term *electric strength* being reserved for values corresponding to electric breakdown. Alternatively, the term *intrinsic electric strength* is used for breakdowns of the electric type, when these are believed to be characteristic of the material under examination. It is, however, by no means easy to obtain values characteristic of any given material, for even when thermal instability and local faults are avoided, spurious values are often obtained owing to ionization of the medium surrounding the test specimen.

Effect of extraneous ionization. This is well illustrated by the experiments of W. Werner[10] who measured the breakdown voltage of thin sheets of mica, rock salt, and sulphur, using as electrodes metal spheres of diameter 1 cm. pressed against the surface of the specimen at opposite points on the two sides. Measurements were made on samples of various thicknesses up to 1 mm. with the sample and electrode system immersed in air, oil, and liquid air. The breakdown voltage was found to vary very greatly with the nature of the surrounding medium, being highest in oil. The apparent electric strength of the material was not independent of the thickness, but the applied voltage V at breakdown was a linear function of the thickness d, i.e. $V=a+bd$. Moreover, the material was never punctured at the point of contact of the electrodes and specimen, where one would expect the field strength to be greatest, but always at a distance of a few mm. from this point. Werner's conclusion was that such breakdowns were originated by subsidiary discharges in the surrounding medium, which were always of the character of a gas discharge, probably occurring in minute bubbles in the case of a liquid. The ions generated by such discharges acquire high velocities in the electric field, and the actual breakdown of the solid he attributed to the impact of such ions on the surface of the solid. Some measurements were made with the electrodes and sample in a high vacuum, gas being removed as far as possible by heating the whole apparatus to 400° C. for many hours during the preliminary evacuation; but even so, the breakdown voltages were much smaller than in oil, which was considered to show that relatively few ions having long mean free paths are capable of initiating breakdown.

The mechanism suggested by Werner for the breakdown of solid dielectrics, viz. fracture of the material by the impact of ions generated in the surrounding medium, is not universally accepted, but it is generally agreed that discharges in the surrounding medium facilitate breakdown, and that apparent values of electric strength obtained in circumstances which permit such discharges do not represent the real electric strength of the material. A good deal of work has therefore been directed towards the explanation and suppression of such discharges. Thus Marx[11] showed that when the medium around the solid under test is a gas, an increase of pressure leads to higher values of breakdown voltage, while Semenoff and Walther[12] showed that when the medium is a liquid, higher breakdown voltages are obtained when the conductivity of the liquid is increased. For example, for some of their work they used as immersion-medium a 12-per-cent solution of acetone in xylene, which may be regarded as a semi-conducting liquid. When using this liquid, central

punctures were obtained and the breakdown field-strength was independent of the thickness of the solid. They therefore concluded that the results were not affected by extraneous ionization.

A. v. Hippel[13] has made a detailed study of such phenomena and has suggested a mechanism for the effect of extraneous ionization which appears to explain most of the observed facts. The electric field between the electrodes will be strongest at points near the two electrode surfaces and as the applied voltage increases the breakdown field-strength of the medium is reached at some such points and brush discharges are set up. The electrons liberated move very much more rapidly than the positive ions, with the result that around the anode surface a positive space-charge is established. If this is uniformly distributed it virtually increases the size of the anode and thus tends to reduce the field-strength below the breakdown value for the medium. At the cathode, on the other hand, the positive space charge left by the moving away of the fast-moving electrons causes an increase of potential gradient, and the brush discharge persists for a time, until eventually the electrons liberated from the cathode produce a negative charge on the surface of the test piece, which is sufficient to stop the discharge. Thus, a uniformly distributed brush discharge should not produce spurious results. But normally the brush discharges, especially at high voltages, are not uniform, but spread from the electrodes in tree-formation, and the branches virtually form additional electrodes. Thus, instead of two symmetrically arranged electrodes, we have in effect an unsymmetrical arrangement consisting of the original anode, and a group of points connected to the original cathode as cathode. The electric field will be strongest in the neighbourhood of the points, and therefore breakdown of the test piece occurs as soon as the field in the neighbourhood of one of the points exceeds the electric strength of the material. A. v. Hippel demonstrated the branched form of the discharges and the difference between anode and cathode, by using as test-pieces photographic film as used for X-ray work, coated on both sides. Both anode and cathode discharges were therefore simultaneously photographed. Somewhat similar methods of recording discharges have been employed by A. Gemant[14, 15] and by A. M. Thomas[16].

Some support is lent to A. v. Hippel's ideas by the following facts: (1) A rapid increase of applied voltage tends to give low breakdown values. The rapid increase tends to produce higher voltages during the time of the first discharges and therefore makes point-action the more probable. A slow rise of voltage favours a more uniform discharge. (2) Undesirable ionization effects occur more readily with alternating than with direct voltage. With alternating voltages the surface charges, which tend to reduce the potential gradients, must be reformed every half-period, at times with change of sign. Thus, greater field-strengths exist and ionization effects are more pronounced. In connexion with (1) it should be remarked that very short "impulse voltages", of a duration of the order of a microsecond, tend to give high breakdown values, the time being insufficient for the establishment of strong subsidiary discharges. In v. Hippel's view the beneficial effect of a semi-conducting medium is due to the fact that the conductivity tends to prevent the

strong local variations of field-strength which must necessarily accompany a branched discharge.

Measurement of intrinsic electric strength. The term *intrinsic* is here used merely to indicate that the value required is independent of thermal instability and extraneous ionization. From v. Hippel's standpoint, extraneous ionization does not affect the electric strength of a solid; it merely falsifies the calculation of the field applied. Thus, the word "intrinsic" becomes superfluous. Nevertheless it has been used, and is convenient if only as an indication that such errors have not been neglected. Thermal instability is easily avoided. One merely has to avoid the use of high temperatures, high frequencies, large thicknesses and materials of high conductivity or power factor. In order to avoid extraneous discharges, it is easy to see that the following conditions are desirable: (*a*) The electrodes should be free from sharp corners, such as would cause a local increase of field strength. (*b*) The space between the surface of the electrode and the surface of the material under test (contact between electrode and test piece will normally be made over a very small area only) should be occupied by material of high dielectric constant, and high breakdown field-strength. This material is in series with the test piece, and therefore the higher its dielectric constant relative to that of the test piece, the lower the field-strength in it as compared with that in the test-piece and the less likely it is to develop discharges. In order to obtain a medium of high breakdown field-strength, gases should be avoided as far as possible, and if present should be at high pressure. Conductivity in the medium may be advantageous, but not if it diminishes the breakdown field-strength.

The arrangement used by v. Hippel[13] is a good example. The specimen is in the form of a thin plate held between two steel electrodes, the contact surfaces being hemispherical and of radius 2 cm. The space between the electrode surfaces and the sample is filled with an easily fusible solid of high dielectric constant. Beeswax, $\epsilon = 4{\cdot}75$, was generally used. The wax was melted and out-gassed in a vacuum, and then allowed to solidify under a pressure of 75 atmospheres round the electrodes and test piece in the form of a tightly fitting ring. The whole electrode arrangement was mounted in an atmosphere of nitrogen, which was maintained during the tests at a pressure of 75 atmospheres, in order to avoid any brush discharge anywhere in the neighbourhood of the electrodes. This mechanical pressure was found not to affect the results. Direct voltage was used and it was increased slowly. A very high series resistance (obtained by the use of a flame or a thermionic valve) was used as a current-limiting device. Occasionally the electrode surfaces were highly polished, and one could then be withdrawn after breakdown without disturbing the position of the specimen, which was left sticking to the other. Microscopic examination of the puncture then showed that it was strictly central. In his earlier experiments, v. Hippel[17] used nitrogen only at a pressure of 75 atmospheres as immersion medium, but it appears that this is not always satisfactory.

The results obtained for rock salt may be quoted as typical. Many specimens were used of thicknesses ranging from about 0·1 to 0·4 mm. The breakdown

voltage was found to be strictly proportional to the thickness of the specimen, giving a constant value of $1{\cdot}5 \times 10^6$ v./cm. for the electric strength. The same result was obtained for natural and artificial crystals, and it is considered that the result is reproducible to better than 10 per cent.

A. M. Thomas[18] has a somewhat different arrangement. He used pure xylene at a pressure of 40 atmospheres as immersion medium. His electrodes were either spheres of diameter 13 mm. or planes of diameter 29 mm. with rounded edges of radius 10 mm. Substantially the same results were obtained for samples tested between two planes, and between a sphere and a plane. He used alternating voltage. Most of his measurements were made with cellulose acetate and mica, so that it is not possible to compare his results with those of v. Hippel. His values of electric strength are not independent of the thickness of the samples which ranged from 0·02 to 0·2 mm. The electric strength diminished linearly with increase of thickness. For thicknesses of 0·04 mm. the values for both mica and cellulose acetate were of the order of 2×10^6 v./cm., while for a thickness of about 0·1 mm. the values were smaller by perhaps 25 per cent. The values are clearly of the same order as those given by v. Hippel. Thomas made many experiments on various immersion media, including semi-conductors such as nitrobenzene, and tricresyl phosphate and insulators such as xylene and transformer oil. He showed that the type of discharge occurring in the medium was of importance, a highly localized discharge being much more harmful than a uniformly distributed one. Pure xylene at 40 atmospheres pressure was preferred as giving the desired type of discharge and having a high electric strength. Impure xylene, although having about the same electric strength, does not appear to give the same form of discharge and is less satisfactory. Mica tested in nitrobenzene at atmospheric pressure gave the same results as in xylene at high pressure. It therefore appears that a semi-conducting liquid of high dielectric constant is satisfactory. Such liquids are, however, active chemically and therefore capable of less general application than xylene.

Breakdown of impulse voltage. Many observations have been made of the breakdown of insulating materials on the application of voltage transients of very short duration, say 1 μsec. or less. These are usually known as impulse voltages. Their duration is controlled by the times of charge and discharge of a condenser in circuits of known resistance, and it is obtained from the calculated time constants of such circuits. A typical impulse generator is shown in Fig. 1.

The two condensers C_1 and C_2 are charged in parallel by means of the high-voltage transformer T and rectifier. When their voltage reaches a certain value V the two equal spark-gaps G_1 and G_2 break down, and the sparks short-circuit the two gaps. Gap G_1 then connects the two charged condensers in series, and gap G_2 allows the series combination (of total voltage 2 v.) to discharge through the specimen under test X. The voltage across the specimen will evidently rise rapidly at a rate largely controlled by the resistance R_1, the inductance of the connecting leads, and the capacity of X, C_1, and C_2 all in series, until the sample also becomes charged to a high voltage, after which the charged system X, C_2, C_1 will discharge through the

resistance R_5 and to varying extents through the other resistances of the network. The maximum voltage across the specimen may be measured by means of a sphere gap G_3. Alternatively, a cathode-ray oscillograph is employed, and the complete voltage-time wave recorded. When the highest possible voltages are required, the number of condensers charged in parallel and discharged in series by means of spark gaps is increased.

The data so obtained are not only of great technical importance as indicating the ability of insulators to withstand lightning discharges and the like, but they also throw some light on the mechanism of breakdown. It is to be noted that the use of impulses of very short duration and small energy almost rules out the possibility of a thermal breakdown. It therefore becomes possible to determine the intrinsic electric strength of materials at temperatures at which thermal breakdown would occur for voltages of long duration. We have also noted that the extreme shortness of an impulse may prevent the establishment of undesirable subsidiary discharges in the surrounding medium, and therefore be advantageous for determinations of intrinsic breakdown strength. However, subsidiary discharges are by no means unknown in work with impulses and the precautions previously discussed should be taken.

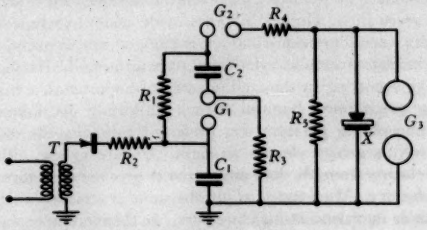


Figure 1. Impulse-generator. $C_1 = C_2 = 0{\cdot}04\ \mu F.$ (say), $R_1 = R_2 = R_3 = 30000\ \Omega.$, $R_4 = 1000\ \Omega$, $R_5 = 10000\ \Omega$.

A good example of a measurement of electric strength by this method is that of K. Moerder[19] on glass. A glass plate of thickness 0·13 mm. was held between copper electrodes (hemisphere against plane), and the whole immersed in glycerine ($\epsilon = 14$). The measured electric strength at 20° C. was $4{\cdot}75 \times 10^8$ V./cm. This is even higher than the values for mica and rock salt quoted above. For materials not easily obtained in the form of very thin plates Moerder, following K. Meyer[20] and Rochow[21], used an arrangement of mercury electrodes. The sample consists of a fairly thick sheet, but a depression made in the middle of one side reduces the thickness at the centre of the sheet to the smaller value required. The sample is then clamped between the ends of two glass tubes, the other ends of the tubes turning up to form a U. The tubes are then filled with mercury in a vacuum, thereby making excellent contact with the sample. Evidently this arrangement should also give freedom from subsidiary discharges. Moerder made many measure-

ments on galalith. He found the intrinsic electric strength to be independent of thickness over the range 0·1 to 0·5 mm., and to be independent of temperature from 20 to 85° C. At higher temperatures the observed breakdown field-strength diminished with rise of temperature. In this region the thermal effects of the discharge are evidently appreciable, but the breakdown value is still higher than that corresponding to purely thermal breakdown which can be observed with voltages applied for a long time. A theoretical discussion of the conditions leading to these partly thermal breakdowns has been given by K. Meyer[20]. As may be expected, thermal phenomena tend to play a more important part in the breakdown the thicker the sample. It is also evident that the magnitude of the thermal effects may be limited by the magnitude of the total energy available in the impulse, which is controlled by the capacitance of the condensers in the impulse circuit. Thus, if measurements of intrinsic electric strength are required, thin samples and impulses obtained with small capacitances are advantageous.

Very high values of electric strength have, however, been obtained even with thick samples. Thus Davis, Standring and Bowdler[27] have obtained a value of $1{\cdot}7 \times 10^8$ v./cm. for glass in thicknesses of 2 and 4 mm. and a value of $1{\cdot}2 \times 10^8$ v./cm. for varnished cloth in thicknesses of 0·2 and 3·7 mm. It therefore appears that thermal effects do not constitute a very serious difficulty in ordinary impulse tests.

By means of measurements with impulse voltages, Inge and Walther[22] have obtained information about the mechanism of the puncture of solid dielectrics when complete breakdown occurs. They used flat plates of glass and rock salt of thickness 1 to 10 mm. with metal electrodes immersed in oil. Subsidiary discharges in the oil were observed, and the plates were examined microscopically at intervals from the time when the subsidiary discharges first appeared until complete breakdown occurred. Signs of discharges within the solid were detected before breakdown occurred. These appeared as branching threads beginning at points on the surface of the solid and terminating in the interior. The threads followed the direction of the electric field in glass but in rock salt they followed paths evidently determined by the crystal structure. The field-strength at which these partial breakdowns first appeared was considered to be characteristic of the material. It was found to be almost independent of temperature for glass over the range -180 to $+100$° C. and to be unaffected by the duration of the applied voltage. The variation with such factors of the voltage required for complete breakdown was attributed to the fact that complete breakdown involves firstly the production of these discharge paths and their extension until they pass right through the material, and then the increase in width of the path as the current passed increases until the material is finally damaged by heat.

It appears probable that the field-strength at which these discharge paths first appear is the intrinsic electric strength of the solid, corresponding to the values obtained by the other experimenters previously mentioned. As, however, subsidiary discharges certainly occurred in the experiments of Inge and Walther, the values of electric strength cannot be calculated from the observed voltage in these experiments, the apparent values of field strength being much too small for reasons

previously discussed. Inge and Walther have, however, also measured values of electric strength of the same order as those quoted above. For the latter experiments they used thin samples immersed in a semi-conducting liquid.

The electric strength of very thin films. In discussions of electric strength some results given by Joffé[23] for very thin materials have long been prominent. Glass was examined in thicknesses down to 0·0001 mm. For thicknesses greater than 0·005 mm. the electric strength was of the order of 10^6 v./cm. but as the thickness diminished beyond this value, there was a rapid increase of electric strength until for a thickness of about 0·0002 mm. a value of the order of 10^8 v./cm. was obtained. A further reduction of thickness caused no further increase of electric strength. About the same maximum value was found for both glass and mica and cooling to the temperature of liquid air did not increase the value. The possibility of obtaining such high values of electric strength evidently raises questions of considerable theoretical interest, and it is curious that these values have apparently not been confirmed by other workers, although it is about ten years since they were first published.

Theoretical considerations. An adequate survey of the theories which have been advanced to explain the breakdown of dielectrics is beyond the scope of this report, but it may be useful to indicate briefly the directions in which explanations have been sought.

The purely thermal character of the breakdown which occurs in certain circumstances has already been indicated in the earlier paragraphs and needs no further consideration. We may here confine our attention to the purely electrical breakdown.

We have seen that values of electric strength of the order of a few million volts per cm. have been reported by several observers using different experimental methods, while Joffé has reported values about a hundred times as great for very thin specimens. Joffé gave an explanation of the lower values in terms of a process of ionization by collision, somewhat analogous to that which is known to occur in gases. The conductivity of many insulating materials is known to increase as the voltage is increased, and it has usually been considered that the conductivity is ionic, and that when the velocity of the ions reaches a certain value they are able to generate new ions, which account for the increased conductivity. Joffé's suggestion was that breakdown occurs when the number of ions per unit volume reaches a certain critical value. It is, however, a very difficult matter to devise any experimental check for the theory. It might however account for the exceptionally high values of electric strength given by Joffé for very thin specimens, for the paths of the ions in such specimens would be so short that the process of ionization might well be severely limited, and breakdown would only occur with higher field-strengths. Joffé suggested that his highest values represented the ultimate electric strength of the material, i.e. the force required to separate its constituent ions and electrons. In this connection it is interesting to consider the equilibrium of a neighbouring positive and negative ion under the influence of electrostatic attraction e^2/r^2, and a force of repulsion $\beta e^2/r^n$. The resultant attraction has a maximum value for a

certain value of r and if the force due to the applied electric field exceeds this value the ions will be torn asunder, i.e. breakdown will occur. If we assume the usual values for the molecular constants, the critical field is found to be of the order of 10^8 v.cm. Rogowski[24] has calculated in a similar way the force required to separate the ions in a regular lattice such as that of rock salt. He obtained a value of the same order.

However, most recent investigators of this problem have taken the view that the conductivity of crystals of insulating material such as rock salt is of an electronic character, and that breakdown is essentially a sudden increase of the electron current. A. H. Wilson's[25] model of a semi-conductor or insulator is that of a perfect lattice in which all the electrons are free. When quantum principles are applied to the problem of electrons moving in the periodic field of the lattice, it appears that the scheme of energy levels consists of bands within which electrons may be found, separated from one another by forbidden areas, which means that electrons cannot exist within the crystal with energy values lying inside these areas. Each band can accommodate a certain maximum number of electrons, and for a given material in a given condition the bands may be completely filled, partly filled, or empty. The effect of an applied electric field is to accelerate the electrons and therefore to increase their energies in continuous steps. This is only possible for electrons which occupy incomplete bands, since only in such bands are there unoccupied levels which the electrons may occupy as their energies are increased. Suppose the number of electrons is just sufficient to fill up the lowest band of allowed energies when all the electrons are in their lowest possible states. Then if a field is applied it will be impossible for any electron to accelerate without making a transition to the second band of allowed energies. This it will be unable to do if the field is small, owing to the finite energy-difference between the two bands. Such a material would be a perfect insulator. It might be realized by a perfect crystal at the absolute zero of temperature. When the temperature is different from zero, there will be a few electrons in the second band and a few unoccupied places in the first band, and some conduction will be possible. As the temperature rises the number of electrons which may give rise to conduction increases and the conductivity therefore tends to increase. The excitation of the thermal vibrations of the lattice will, however, tend to decrease it. At low temperatures the first effect must predominate and therefore we have the rapid rise of conductivity with temperature which is usually observed with insulating materials. At higher temperatures the effects of the thermal vibrations may predominate and the temperature coefficient of resistance may change sign—an effect which is observed for semi-conductors.

In order to account for electrical breakdown we have therefore to discover the existence of a critical field-strength, such that as the applied field passes this value, the number of electrons in an unfilled band suddenly increases. A. v. Hippel[13] has suggested that the critical field-strength is that which raises the electron energies to the value corresponding to the residual rays of the material, i.e. to the value $h\nu_m$, where h is Planck's constant and ν_m is the limiting frequency of the

thermal spectrum. Electrons with an energy exceeding this value are considered to be unable to lose energy in producing thermal agitation by collision with ions of the lattice. Their energies therefore increase until they are sufficient to cause rupture of the lattice by collision.

Zener[26], from a consideration of the effects of an applied electric field on the bands of energy levels, has calculated the rate at which electrons may escape from the lower into the upper energy bands under the influence of such a field. He concludes that it is possible to account in this way for a very sudden increase of electron conduction current, such as would constitute breakdown, at values of field strength of the order of 10^6 v./cm.

Whitehead[5] has suggested a possible broadening of the bands under the influence of the applied field, breakdown occurring as soon as neighbouring bands overlap.

The available experimental evidence is not adequate to establish any existing theory. Data for about a dozen alkali halides have been given by v. Hippel, from which it appears than an explanation on the lines suggested is at least plausible. We have, however, not yet reached the stage at which different observers have given the same results for several representative materials. The more recent values do, however, appear to be characteristic of the materials under investigation and not merely peculiar to the technique employed so that there is good reason to hope that further work in this direction will be profitable.

References

(1) Wagner, K. W. *Trans. Amer. Inst. elect. Engrs*, **41**, 288 (1922).
(2) Hartshorn, L. and Ward, W. H. *J. Instn elect. Engrs*, **77**, 723 (1935).
(3) Hayden, J. L. R. and Steinmetz, C. P. *Elect. World, Lond.* (1922).
(4) Schumann, W. O. *Z. tech. Phys.* **6**, 439 (1925).
(5) Whitehead, S. *World Pwr*, **26**, 72 (1936).
(6) Fock, V. *Arch. Elektrotech.* **19**, 71 (1927).
(7) Whitehead, S. *Dielectric Phenomena*, 3, 299 (1932), London: Ernest Benn Ltd.
(8) Mason, H. E. *Trans. Amer. electrochem. Soc.* **65**, 73 (1934).
(9) Whitehead, S. and Nethercot, W. *Proc. phys. Soc.* **47**, 974 (1935).
(10) Werner, W. *Ann. Phys., Lpz.*, **86**, 95 (1928).
(11) Marx, E. *Elektrotech. Z.* **50**, 41 (1929).
(12) Semenoff, N. and Walther, A. *Die Physikal. Gundlagen d. Elektr. Festigkeitslehre*, p. 121 (1928), Berlin: Springer.
(13) V. Hippel, A. *Z. Phys.* **75**, 145 (1932).
(14) Gemant, A. *Z. tech. Phys.* **12**, 250 (1931).
(15) Gemant, A. *World Pwr*, **24**, 8 (1935).
(16) Thomas, A. M. *World Pwr*, **24**, 86 (1935).
(17) V. Hippel, A. *Z. Phys.* **67**, 707 (1931).
(18) Thomas, A. M. *World Pwr*, **24**, 125 (1935).
(19) Moerder, K. *Arch. Elektrotech.* **24**, 174 (1930).
(20) Meyer, K. *Arch. Elektrotech.* **24**, 151 (1930).
(21) Rochow, H. *Arch. Elektrotech.* **14**, 361 (1925).
(22) Inge, L. and Walther, A. *Arch. Elektrotech.* **24**, 259 (1930).
(23) Joffé, A. *Physics of Crystals* (1928), McGraw, Hill.
(24) Rogowski, W. *Arch. Elektrotech.* **18**, 123 (1927).

(25) Wilson, A. H. *Proc. roy. Soc.* A, **133**, 271 (1931).
(26) Zener, C. *Proc. roy. Soc.* A, **145**, 523 (1934).
(27) Davis, R., Standring, W. G. and Bowdler, G. W. Conférence Internationale des Grands Réseaux Électriques à Haute Tension. Paris, 1937, Paper No. 311.

§ 2. MEASUREMENT OF CURRENT AT RADIO FREQUENCIES

By T. Iorwerth Jones, M.Sc., A.R.C.S., D.I.C.

The main difference between the behaviour of a low-frequency alternating current flowing in an electrical circuit and a high-frequency current is that the former confines itself almost entirely to certain prescribed conducting paths while the proportion of the latter which restricts itself to those paths becomes less and less as the frequency is raised. This implies that the current choosing the conducting path varies from point to point, with the result that the current recorded by an instrument will depend upon the point at which that instrument is introduced. Its readings will, therefore, rarely represent the total current flowing in a given circuit or be a clue to the total power dissipated.

It will be evident that a condition to be satisfied at very high frequencies is that an instrument should be so small and compact that the current at all points in it is substantially the same.

How difficult of realization some of these conditions are may be gathered from the fact that no claims are made that the behaviour of instruments hitherto commercially produced is capable of accurate prediction at frequencies exceeding 100 Mc./sec. This can be easily understood when it is borne in mind that the effects referred to above increase with frequency raised to the second power. It is therefore safe to say that a condition of resonance is reached in most commercial instruments between 200 and 300 Mc./sec. In the case of some laboratory measuring devices, these frequencies may be exceeded but only to a very small extent. Beyond 100 Mc./sec. most instruments become mere indicators which can serve approximately for comparison purposes.

With modern apparatus continuous oscillations of frequencies of several hundred Mc./sec. are readily produced, and frequencies of 50 and 1500 Mc./sec. are employed commercially for television and telephony respectively. The measurement of current at high radio-frequencies has therefore become of considerable interest.

Every type of instrument hitherto described essays to measure the current traversing a conducting path. Problems of frequency dependence of current-distribution are therefore common to them all. The property of an electric current which has found most general application has been its heating effect. Current is passed through a conductor, and one effect or another of its rise in temperature is observed. The heating effect, however, is dependent upon the frequency of the alternating current due to the so-called skin effect. It therefore follows that it takes a smaller current at a radio-frequency to produce the same heating effect in a wire than at power frequency or with direct current. The current, I_H, passing at a high

frequency for which the resistance of the heater is R, bears the following relation to the direct current I_0, for which the resistance is R_0:

$$I_H/I_0 = (R_0/R)^{\frac{1}{2}}.$$

If the resistance of the heater is not to change considerably with frequency, its thickness must not exceed the equivalent penetration of the current at the frequency in question. This restricts the transverse dimensions of the heater to the very smallest that will carry the current without excessive heating and reduces the overload factor of the instrument.

But this is not the only handicap with which the designer of high-frequency ammeters has to contend. The heater possesses both inductance and resistance, which are further associated with capacitance, part of which is the capacitance across the terminals of the instrument. In its simplest form the circuit of an ammeter is represented by figure I, in which L and R represent the inductance and resistance, respectively, of the conducting path of the instrument.

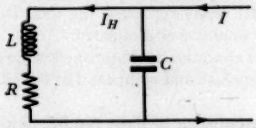


Figure 1. Equivalent circuit of an ammeter.

It may be shown that the current I_H flowing through the heater is given by the expression

$$\frac{I_H^2}{I^2} = \frac{1}{(1 - LC\omega^2)^2 + R^2C^2\omega^2},$$

or, if the frequency is sufficiently below the resonant frequency of the heater circuit

$$I_H/I \doteqdot \frac{1}{1 - LC\omega^2},$$

so that the complete relationship between the high-frequency current, I, entering the instrument and its direct current or power-frequency equivalent, I_0, is

$$I_0/I \doteqdot \frac{1}{1 - LC\omega^2}(R/R_0)^{\frac{1}{2}}.$$

In an ideal instrument L and C should be capable of exact definition or, to say the least, it should be possible to determine the first correction term involving them with certainty. The number of actual instruments in which this can be done are few, for, in general, devices are introduced into close proximity with the heater which possess capacitance to the heater and to earth, respectively. The former is difficult to allow for, and the latter may depend upon the location of the instrument and upon its potential relative to earth. Besides these complications, every instrument is liable to be influenced by the capacitance between the leads connected to its terminals. The reading is therefore subject to a number of undesirable effects

which cannot be calculated or measured. Of these, capacitance is not the only factor, for the unsymmetrical disposition of metallic objects, such as thermo-junctions, will tend to affect the symmetrical distribution of the current in the heater. Any unsymmetrical distribution of current increases the resistance of the heater.

One other handicap is encountered due to the necessity for cutting down the thickness and therefore the current-carrying capacity of the individual heaters, which implies that a number of elements in parallel are necessary if the total current capacity is to be maintained. At power frequencies the current distribution between a number of elements will be solely determined by their resistances. It is therefore possible to use external shunts. At high frequencies, the reactances will determine the distribution, and the designer is confronted with the problem of securing perfect symmetry, not only as regards resistance but also as regards reactance, between the members and the incoming and outgoing leads. For this reason the use of external shunts is not permissible. The shunting factor will vary with frequency unless the elements are identical and symmetrically disposed to one another and to the external circuit.

Experimenters have resorted to various means of registering the heating effect. The steps which have been taken in recent years have possessed few novel features; on the contrary, older projects have been revived and modified to meet the conditions encountered with the upward extension of the frequency range.

Thus, the air-thermometer ammeter which to-day is accepted as the most satisfactory method, on account of its sensitivity and simplicity, of measuring very small currents, was first utilized by Snow Harris[1] in 1827, for observing the relative powers of various metallic substances as conductors of electricity.

The methods utilizing the heating effect of a current traversing a conductor will be subdivided into groups, which will be designated by headings conveying the essential nature of the process of observation.

(*a*) *Dilatometer ammeters.* Adopting the air-thermometer idea of Snow Harris, Espinosa de los Monteros[2] enclosed the bulb in a vacuum jacket and observed the expansion of the air by means of a liquid index in a capillary. To obviate the uncertainty in the position of the index due to a drift in the external temperature, Scheibe[3] utilized two gas reservoirs, disposed symmetrically in outer jackets inside a cork enclosure and communicating with one another through a capillary. The index now indicated the expansion of the air in the heated bulb relative to that of the air in the other. Schmitz[4], however, constructed a fully compensated system by inserting a conductor into both bulbs. Radio-frequency current traversed one heater, while the direct current flowing along the other heater in the second bulb was adjusted so that a small liquid index in a communicating tube remained at its initial position. The index was observed under a microscope. In this manner Schmitz measured currents requiring an expenditure of power of the order of only 10^{-6} w. with an accuracy of 5 per cent. Takaya[34] also used a similar differential air-thermometer as a standard for testing thermal instruments over a range 5 to 500 ma. at frequencies as high as 10 Mc./sec. Readings obtained with direct current repeated to 0·6 per cent.

The air-thermometer ammeter suffers mainly from the disadvantage that the heater has to be fairly long and the bulb somewhat large. Against this it possesses the advantage that no auxiliary equipment is necessary for observation purposes. The circuit of the instrument is simple and isolated in space.

Herschberger[5] dispensed with air as calorimetric substance and substituted water in a vacuum jacket. He observed the rise in temperature of the water in a given time. By this means he was able to increase the current-carrying capacity of the heater and so reduce the skin effect. Wilmotte[6], on the other hand, utilized his thermometric substance as heater and resorted to a device which resembled a mercury thermometer. The current was passed through the mercury in a fine tube, and its rate of expansion was observed in a capillary of very fine bore. This instrument possessed a higher inductance than most types and used an unusually long heater. Both of these features are disadvantageous at high radio-frequencies.

(*b*) *Extensometer ammeters.* These instruments are commonly described as hot-wire ammeters. They rely upon an observation of the thermal expansion of the heater itself. The principle and the mechanism employed have found very general application at power frequencies for a considerable time, but apart from mechanical considerations no restrictions had been placed on the cross-section of the elements; in fact, the cross-section needed to be small if the necessary flexibility was to be obtained, but at power frequencies external shunts could be employed. Moreover, the use of such shunts at high frequencies is not permissible. Broca[7], therefore, conceived a system of current subdivision which maintained an exact symmetry of current distribution by disposing the heaters like the wires of a cage on a cylindrical surface and leading the current in and out at the centre of the end-surfaces. In view of the symmetry he limited his observations to the sag of one wire only. At the frequencies involved in those days the disymmetry resulting from this procedure had no serious results. Taylor[8], on the other hand, attempted to incorporate a compensating element. A wire carrying the high-frequency current and another carrying direct current were connected through small springs with the pointer of the instrument. When the sags in both wires were equal, the pointer registered the same initial position as it occupied when no current flowed in either wire. The direct current was adjusted for this equality, and the relation existing between the two currents was known from previous observations with direct current traversing both wires.

As far as can be gathered from the literature, the method of compensation has not been widely used. Schlesinger[9], who has specialized in the application of the hot-wire method for the measurement of small currents at high frequencies, has resorted to a very ingenious device for the calibration and identification of the current flowing. He employs a bulb 8 cm. long in which a fine metallic string 4 cm. long is maintained in tension. The string is excited mechanically through the bulb by a vibrating diaphragm which is actuated from a valve oscillator producing an adjustable audio-frequency. With the flow of current through the string its resonant frequency changes, and the current flowing can be identified with the frequency to which the string responds. The device is very sensitive and permits of the

detection of currents which are but a tenth of that required to cause the wire to glow. There is, therefore, little danger from overloading, and the power consumption is very low. Schlesinger claims that his device is suitable for the measurement of currents of less than 100 ma. at frequencies of the order of 100 Mc./sec. Observation of the string is made under a microscope or by projection.

Borgars[10] has observed the sag in the heated wire directly. The wire is stretched excentrically in a small glass bulb and carries a small weight attached to it at the centre. This weight rests normally on the wall of the bulb, but during observations, after the bulb has been rotated through a small angle, it hangs freely, and its movement with the passage of the current is observed in a microscope.

Of all the hot-wire instruments the one which has been worked out in most detail is that described by Fortescue[11]. It is intended for comparatively large currents, and models carrying currents ranging from 0·25 to 4 amp. have been described.

Each heating unit consists of a fine wire of platinum-silver mounted coaxially in a metal tube, which serves as a return path, and is connected to it at one end through a flexible strip to which the recording system is attached. This construction, besides providing adequate screening and the absence of appreciable mutual inductance between external circuits and each heating unit, of which there may be several in parallel, facilitates the task of the precise calculation of the frequency errors associated with the instrument. The circuit lends itself easily to the determination of the effect of distributed capacitance and inductance by the application of transmission line theory. The effect of terminal capacitance is also calculable. Outside this simple circuit there is no extraneous capacitance, within the instrument, to be considered, for the only attachment to the hot-wire is at its junction with the screen, i.e. the direct expansion of the wire, and not the sag, is effective in operating the recording mechanism. These corrections, however, form only about a quarter, at 100 Mc./sec., of the total correction. In a case quoted by Fortescue, the current recorded at 100 Mc./sec. exceeds the d.-c. equivalent by a factor 1·417. Skin effect is responsible for a factor 1·295. (The wires range from 0·008 to 0·015 cm. in diameter.) This leaves 1·093 as the factor contributed by distributed and terminal capacitance. It is clear that an accuracy of 1 per cent in the final determination will require considerable uniformity throughout each wire and between the units, as well as an accurate knowledge of their dimensions. The presence of the screen has contributed materially to the difficulty in reducing the skin effect correction, for it has restricted the heat-dissipation from the wires. Granting that the estimated corrections may not be exact to within 1 per cent, at 100 Mc./sec., the instrument has, nevertheless, a wide range of application. At 10 Mc./sec. the total correction is only 1 per cent, while at 50 Mc./sec. the total correction is 14·5 per cent, of which only 1·8 per cent is contributed by capacitance effects.

Fortescue has succeeded of late in reducing the capacitance corrections by shortening the length of the hot wires from 7·5 to 5 cm., thus producing instruments whose external dimensions were reduced from 20 to 12·5 cm.

Single units have been made to carry a full load 0·25, 0·5, and 1 amp. Higher ranges of current are obtained by multiplying the number of units. Fortescue has endeavoured to preserve the utmost possible symmetry in their disposition with respect to the in-going and out-going leads, so as to secure equality of current distribution between them. Should any inequality in expansion arise, however, he has adopted a system which averages out the expansion of all the units and which thus compensates to some extent for imperfections in construction. There cannot be a doubt but that these instruments represent the peak yet reached in the design of hot-wire ammeters for the measurement of moderately large currents at radio-frequencies. The accuracy of the calculated corrections was checked by experimental comparison of models differing somewhat amongst themselves in construction.

(*c*) *Thermo-electric ammeters.* Under this heading it is proposed to deal with thermo-ammeters which rely on heaters and thermo-junctions for the linking up of the heating effect in a conductor with the observations of a d.-c. instrument. Of thermal instruments these have been the ones most generally employed since the days of Duddell. The departures in design have been few apart from the substitution of platinum-silver and platinum-iridium for the constantan of the early days, and of thin strip for the fine wires originally used. Strips show a larger skin effect at first, but they possess a definite advantage over wires at higher frequencies. Of recent years the corrections involved have been considerably reduced by the appearance on the market of thin-walled tubes of the alloys referred to above. It is now possible to obtain tubes 0·07 cm. in diameter with a 0·0025-cm. wall[13], capable of carrying 5 amp. in air, which show an increase in resistance over the d.-c. value of not more than 2 per cent at a frequency of 80 Mc./sec. This accomplishment also assists in the reduction of the impedance correction, since the inductance of the larger tube is considerably less than that of the solid wire of the same current-carrying capacity and the tube can be increased indefinitely in size for still larger currents without involving any increase in skin effect.

Papers by Goodwin[12] and Miller[13] of the Weston Electrical Instrument Corporation have recently indicated that the latest developments in the design of thermo-ammeters have been in the direction of shortening and reducing the cross-sectional area of the heaters. This inevitably reduces the temperature-rise in the element due to increased conduction away from it, but sensitivity is not usually lacking and it may be compensated by recourse to finer heaters. Thus a 0·028-cm. heater wire about 0·6 cm. long can be replaced by a 0·008-cm. wire 0·075 cm. long. The temperature rise is the same and the overload capacity remains unchanged. This results in a reduction of the impedance effect and the skin-effect correction.

One feature from which thermo-ammeters have always suffered has been the shift in zero after the current has been flowing for a time. Two attempts have been described for coping with this disadvantage. Goodwin advocates the placing of a second conductor, identical with the heater, in thermal but not in electrical contact with the end-pieces. The ends of this conductor are separated by very thin strips of mica from the end-pieces. Two junctions are then employed in close contact

with the centres of the two conductors and the galvanometer then records the temperature difference between the two which is independent of the rise in temperature of the whole instrument. Coleman[14] has suggested a slightly different system which also uses a second thermo-junction. In this scheme the latter is in thermal contact with one of the heater supports. In these two ways the troublesome zero drift is largely compensated.

Thermo-ammeters, in general, are associated with much auxiliary apparatus comprising junctions, unsymmetrically disposed, in close proximity to the high-frequency circuit and direct-reading d.-c. instruments built into the same envelope and unscreened. This arrangement does not offer the same opportunity for precise calculation of the constants of the circuit as, it was seen, was offered by the Fortescue hot-wire ammeter.

Nergaard[15], however, has employed thermo-junctions for the measurement of the power supplied by short-wave oscillators operating on wave-lengths below 2 m. They covered a power range from 0·1 mW. to 50 W. In their construction he departed from the lines followed by the designers of commercial thermo-junctions who have adhered very closely to the pattern of the standard thermionic receiving valve, even to the extent of including unnecessarily long, adjacent leads terminating in a pin and socket mounting. The latter, from a high-frequency point of view, are impedimenta which contribute both inductance and capacitance to the apparatus. Nergaard, therefore, copied the Acorn valve-mounting, which is of the simplest, keeping the inductance of the leads and the capacitance across the heater circuit down to a minimum. This step represents a distinct improvement in heater and thermo-junction design, for nothing could be more compact.

(*d*) *Optical ammeters.* In previous sections it has been emphasized that the tendency in the design of high-frequency ammeters is to resort to the smallest cross-section in the heater which will carry the current which has to be measured, in order that the skin effect shall be kept as low as possible. As a result, the main objection to thermal instruments of the hot-wire type described in the two preceding sections has been their small overload capacity. This feature has rendered them particularly vulnerable in high-frequency work where the value of the current is liable to sudden fluctuations due to unforeseen changes in the factors determining the resonant condition of the circuit. A given value of current is frequently realized by approach towards resonance in a tuned circuit. Occasionally, conditions over which the experimenter has no control modify the resonant condition sufficiently to cause a sudden rise in current against the ill-effects of which the instruments are insufficiently protected. In the optical ammeter this danger is even greater, for the whole principle involved is the loading of a heater to a point at which it gives off visible radiation, the intensity of which can be measured and correlated with heating current. For this reason this type of instrument has no application outside the laboratory, where it may serve, nevertheless, as a valuable standard for calibration purposes. In this type of instrument the skin effect must be a minimum for a given current-carrying capacity, since the dimensions of the heaters are a minimum and because the method calls for little auxiliary apparatus in close

proximity to the heater and therefore liable to disturb the uniform distribution of the current. It has to be borne in mind, however, that in most of these so-called photo-ammeters the heater is enclosed in an evacuated bulb. Under those conditions its current-carrying capacity is considerably reduced. Stanek[16], for example, gives an instance in which a certain element carries 5 amp. in air, while its capacity *in vacuo* is as low as 2 amp. Immersed in a liquid, the current-carrying capacity of the same heater would be considerably greater than 5 amp.

This type, again, made its first appearance as long ago as 1904, when Orlich[17] of the Physikalisch-Technische Reichsanstalt used a Holborn and Kurlbaum optical pyrometer for observing the luminous radiation emitted from wide strips of platinum for the smaller alternating currents, and from hollow tubes for the larger currents of the order of several hundred amperes. Steinhaus[18] and Clayeux[19] used photometric methods of observation, while Jouaust[20] employed a Féry pyrometer in 1922 for the calibration of high-frequency ammeters. Later Pession and Gorio[21] adopted a tungsten filament and utilized the photo-electric cell. Their measurements covered a current range of 0 to 5 amp. and were carried out at frequencies extending from 250 kc./sec. to 7 Mc./sec. The filament was chosen according to the skin effect which could be tolerated. Very little has been written of recent years about the technique employed, but the literature abounds in proof that the photo-ammeter is being increasingly employed as a standard for the calibration of other ammeters at radio frequencies. From the writings of Stanek[16] it is clear that progress is being made along the same lines as characterized the latest developments in thermo-ammeters. The impedance of the heater is being reduced both by curtailing its length and by increasing its cross-section and substituting tubes for solid conductors. This process is developed still further by operating the heaters in air instead of *in vacuo*. Such a course can only be adopted with platinum which, unfortunately, is a metal with a large temperature-coefficient of resistance. This factor introduces a little uncertainty if the dimensions of the heater are such as require an accurate calculation of the skin effect.

Finally, it should be noted that the photo-ammeter can be freed from the uncertainties which surround the use of the thermo-ammeter. Its circuit can be far more exactly defined. It is simple and should be amenable to calculation. For this reason it offers great advantages at very high frequencies.

(*e*) *Thermionic ammeters.* As the name suggests, this class of instrument relies upon the correlation of the emission current from a filament with the heater current which produces it. In its original form as used by Hazel[22] a simple comparison was made between the direct current passed through the filament which gave the same emission as the high-frequency current which it was desired to measure. In this form any current could be measured for which a suitable filament could be found, but the filament had to be enclosed *in vacuo*. If filaments of pure tungsten were employed the method had no advantage to offer over the photo-ammeter since the circuit was encumbered with filament and anode-batteries and a d.-c. milliammeter. Its overload factor was also low. For large currents, therefore, it could not be recommended.

For currents of a few milliamperes, however, it has fulfilled a want in a modified form and in this form it has also scored the advantage that its overload factor is much higher than that of the ordinary thermal instrument.

Martyn[23] converted a triode into a diode by connecting grid and anode together and superimposed the small high-frequency current to be measured upon a larger high-frequency current of the same frequency derived from an auxiliary source. The filament was activated by the latter, and the change in emission due to the superposition of the smaller current was observed. By securing exact synchronism in frequency the two currents were either exactly in phase or opposite in phase. The change in emission current was therefore related to the peak value of the high-frequency current, and the correlation between them could be deduced from graphical measurements of the characteristics of the valve employed.

In 1935, a number of papers appeared by Barlow[24] describing a commercial multi-range instrument incorporating the same idea but purporting to measure the root-mean-square value of the high-frequency current. Barlow superposed the high-frequency current upon a direct current and claimed that a current of from 2 or 3 to 30 ma. could be measured with an error of less than 1 per cent throughout a frequency range extending from 25 to 5×10^6 c./sec. The response of the instrument falls off outside this frequency range. The whole apparatus, with the exception of the accumulator battery, is contained in a portable box. A bridge circuit is employed. Two arms consist of resistances while the other two comprise the anode-filament impedance of two almost identical valves. The same direct current flows through the filaments of both valves whereas the high-frequency current traverses one only. A unipivot galvanometer is connected in series with one of four resistances, according to the current range required, across the bridge and records the out-of-balance current. Provided that balance is established at the outset with only direct current flowing through the filaments, the galvanometer can be calibrated to read off the high-frequency current directly. Chokes and condensers are necessary for isolating the high-frequency and the d.c. circuits and two variable resistances for obtaining the initial balanced condition.

The instrument is quick in response, provides four ranges, has a low impedance of 35 Ω., and possesses a high overload capacity. Its range, however, is rather limited in comparison with the air-thermometer ammeter of a similar current rating.

Current transformers. Since the skin-effect limits the size of the conductor which can be used in a thermal instrument, it has been found convenient to use current transformers which reduce the current which has actually to be measured to a much lower value. With the aid of such intermediaries thermo-junctions carrying well under 1 amp. can be used for measuring primary currents of tens of amperes. If M represents the mutual inductance between the primary and the secondary windings, and L and R refer to the secondary circuit, in which the instrument is placed, the ratio of transformation is given by

$$\frac{M\omega}{\sqrt{(L^2\omega^2 + R^2)}},$$

in the region of frequency where capacitance effects are negligible. At frequencies below 1 Mc./sec., Campbell and Dye[25] increased the reactance $L\omega$ in proportion to the resistance R of the secondary circuit by constructing the toroidal transformers on a laminated iron core. This represented a great improvement over air-core transformers and the ratio approximated very closely to the figure M/L over the whole range examined.

At higher frequencies losses in the iron core make a large contribution to the effective resistance of the secondary and spoil the advantages which the use of iron possessed at the lower frequencies. The core also increases the capacitances associated with the windings. Reversion to an air-core transformer results in the ideal ratio M/L being only approached at comparatively high frequencies, since R is predominantly the resistance of the instrument, and the ratio can only be determined experimentally at high frequencies, where $L\omega$ is large in proportion to R. It should be determined at the lowest frequency at which R is negligible in comparison with $L\omega$, and the value so obtained may then be used at higher frequencies. But at those higher frequencies, capacitance effects in the transformer enter into the problem and these are not very amenable to calculation.

Albrandt[26], however, has devised a transformer system which he used for measuring currents of from 5 to 50 amp. at frequencies as high as 15 Mc./sec. He defined the main current circuit accurately by giving it a rectangular configuration and located another smaller rectangular circuit in close proximity to it. The mutual inductance of the two circuits was calculated and the self-inductance of the measuring circuit. Albrandt claimed an accuracy of about 2 per cent between 20 and 150 m.

Rectifier ammeters. The types of instruments described hitherto have relied without exception upon the heating effect of a current for its determination. Other properties of currents have been employed. The place of a thermo-junction is often taken in the secondary circuit of a current transformer by a rectifier instrument. Considerable improvements have been effected in recent years in providing rectifiers the characteristics of which are not too different at radio frequencies from what they are at power frequencies. Owing to their finite cross-sectional area and their small thickness, they possess a finite capacitance which renders their impedance a function of frequency. This does not rule out their utilization in radio-frequency instruments but it does mean that the calibration of the latter at radio-frequencies bears so distant a relation with that at power frequencies that they have to be specially calibrated in the belt of frequency for which they are intended. Such instruments are made by Salford Electrical Instruments, Ltd., and calibrated for a given frequency range. This class of instrument possesses a fairly high overload factor.

Cathode-ray ammeters. Within the last few years it has been suggested that the current to be measured should be passed through a suitable coil and caused to deflect a beam of cathode rays. Matthias, Borries and Ruska[27] have described the construction of such a coil which is estimated to have a self-inductance of 10^{-8} H. The apparatus may be calibrated at a low frequency. It has the advantage of a

linear scale connecting current and deflection. They claim to have measured currents of the order of 10 amp. at a frequency of 1 Mc./sec. Brüche[28] has also given a brief description of a similar arrangement used at the A.E.G. Research Institute.

Dynamometer ammeters. Most of the electrodynamic instruments which have been described suffer from the defect that their calibration is an uncertain function of frequency. This must apply to instruments which depend on the deflection of discs or closed coils in which oscillations are induced, such as those described by Pierce[29] and Keinath[30] and to a lesser extent to a type developed by Turner and Michel[31] in recent years for use at frequencies between 1 and 100 Mc./sec. The instrument consists of two single-turn coils of 3·5 and 2 cm. diameter respectively. The smaller coil is suspended by a quartz suspension in the field of the larger coil. When current flows through the primary, the smaller coil tends to set itself in the position of zero coupling but its inertia carries it beyond this position and the coil oscillates about the position of zero coupling. The mechanical control is negligible in comparison with the torque exercised by the electromagnetic field which is proportional to the square of the current and consequently the frequency of the oscillations set up is determined by the current and the constants of the circuit; this frequency is proportional to the current flowing in the primary circuit.

The instrument can be considered as a current transformer which does not need an auxiliary current measurer in the secondary circuit. For this reason the preponderance of reactance over resistance in that circuit, necessary so that the current ratio may be independent of frequency, has been reached at a frequency of about 1 Mc./sec. This implies that the instrument may be calibrated against a thermojunction at that frequency and that the calibration will hold at still higher frequencies until capacitance effects intervene. The simplicity of the two windings, however, minimizes the effects of capacitance and extends its useful range of frequency. It is unfortunate that the authors do not indicate the current range for which the instrument has been used, but presumably the instrument was intended for large currents only. The power consumed, however, is relatively small since the main restriction on the size of the conductors is the weight of the elements, especially the secondary ring. Protection from air currents is afforded by enclosing the instrument in a glass tube, the diameter of which is chosen to make the damping of the oscillating system critical.

Moullin[32] has endeavoured to produce an ammeter of a form which can carry, unshunted, currents up to several hundred amperes and in which the frequency error can always be calculated. Two models have been described capable of measuring currents of 2 and 25 amp., respectively. They do not differ in any essential detail except as regards the flexibility of the strips supporting the movable part. In both, the repulsive force between two parallel cylinders arranged in series carrying current in opposite directions is deduced from observations of the movement of a pointer attached to the cylinder which is free to move. The displacement is practically proportional to (the root-mean-square value of the current)^2 and inversely proportional to the distance between the centres of the cylinders. One cylinder is of

solid copper and is fixed in position; the other is an aluminium tube, for lightness, and is carried by two flexible strips, which supply the mechanical control and which convey the current to and from the tube. The initial position of the tube can be varied by rotating the instrument about a horizontal axis and so varying the gravity component of the controlling force. Its correct adjustment may be made to a pre-arranged position by observing the tube through the microscope provided for viewing the quartz pointer of the instrument. The cylinders are then at their correct initial distance apart. Current is then passed and the excursion of the pointer from its zero position is observed in the microscope. The maximum movement is 1 cm. which can be read to rather better than 1 per cent at full deflection. The instrument is calibrated with direct current and at this stage the vertical component of the earth's magnetic field has to be neutralized by accurately locating a magnet provided with the instrument.

In an actual instrument the cylinders are 10 cm. long, the strips supporting the movable tube 3 cm., the diameter of the cylinders 1·1 mm., their distance apart 4 mm., the maximum excursion of the moving cylinder 0·2 mm. The cylinders are supported at the centre of a screening, copper enclosure 190 mm. long, 78 mm. in diameter and 3 mm. thick, which is filled with oil. The weight of the aluminium tube is taken by the oil, which also assists in cooling the suspension strips and in damping the motion of the cylinder—thus rendering what would otherwise be a fragile instrument robust and portable. Apart from the microscope, the instrument takes up considerable table space being 35 cm. long, 17 cm. wide, and 24 cm. high.

Correction factors have been worked out for the uneven distribution of the current in each cylinder due to the presence of the other, and of the screen (*a*) round the periphery of the cylinders and (*b*) along their length. The eddy-current effect of the screen can be ignored if the instrument is initially calibrated at a frequency exceeding 125 c./sec. In that case the order of the corrections for the 2 amp. instrument, neglecting the effect of terminal capacitance, is as shown in table 1.

Table 1

	Frequency				
	5×10^6	10^7	10^7	2×10^7	3×10^7
Correcting factors (*a*)	1·025	1·035	1·04	1·04	1·04
(*b*)	1·000	1·000	1·006	1·024	1·054
Total	1·025	1·035	1·046	1·064	1·094

This instrument absorbs little power. For a current of 1·5 amp. it absorbs only 0·7 w. The larger instrument absorbs 6 w. for 21 amp. with direct current and 28 w. at a frequency of 3 Mc./sec. Owing to the bifilar construction, the instrument has practically no external field even without the screen; with it the instrument is completely protected against induction from outside.

Moullin has outlined in his published papers other possible developments for coping with larger currents. He envisaged an instrument with a small cylinder fixed within a movable, and much larger tube. The existing instruments, though

cumbrous, may offer a useful standard of comparison with other instruments of a similar range.

Comparison of ammeters. The last word must rest with experiment and to this end it is necessary to evolve measuring technique as well as instrument technique. If two instruments are connected in series in the same circuit, the essential condition which has to be satisfied is that the same current flows internally between the two terminals of one instrument as flows in the other. Then, and only then, is the comparison valid. Fortescue and Moxon[33] have proposed what is probably the most satisfactory method of observing this condition. They set up a bridge circuit. Two variable condensers occupied two arms in series across the inducing coil. The two ammeters formed the other two. A valve voltmeter was connected between the earthed common point of the two ammeters. The value of current required was obtained by adjusting the capacitance of the two condensers simultaneously, sharing the capacitance between the two until the valve voltmeter showed a minimum deflection. The common point of the ammeters was then at earth potential and therefore no current passed between this point and any neighbouring earth-connected objects. Better results were obtained by connecting together the terminals of the instruments which were not directly connected to the screens.

References

(1) Snow Harris, W. *Philos. Trans.* **18** (1827).
(2) Espinosa de los Monteros, A. *Jb. drahtl. Telegr.* **1**, 323 (1907).
(3) Scheibe, A. *Z. Hochfrequenztech.* **25**, 12 (1925).
(4) Schmitz, W. *Z. Hochfrequenztech.* **27**, 18 (1926).
(5) Herschberger, D. C. *Elect. J.* **14**, 59 (1917).
(6) Wilmotte, R. M. *Nature, Lond.*, **124**, 651 (1929).
(7) Broca, A. *Bull. Soc. int. Élect.* **9**, 423 (1909).
(8) Taylor, A. H. *Phys. Rev.* **34**, 365 (1912).
(9) Schlesinger, K. *Z. Hochfrequenztech.* **36**, 62 (1930).
(10) Borgars, S. J. *Wireless Engr*, **11**, 134 (1934).
(11) Fortescue, C. L. *J. Instn elect. Engrs*, **68**, 556 (1930); **79**, 179; 200 (1936).
(12) Goodwin, W. N. *Elect. Engr*, **55**, 23 (1936).
(13) Miller, J. H. *Proc. Inst. Radio Engrs, N.Y.*, **24**, 1567 (1936).
(14) Edgecumbe and Ockenden, F. E. J. *Industrial Electrical Measuring Instruments*, p. 201 (1933). London.
(15) Nergaard, L. S. *Proc. Inst. Radio Engrs*, **24**, 1207 (1936).
(16) Stanek, J. *Arch. tech. Messen*, **4**, T66–67 (1935).
(17) Orlich, E. *Z. InstrumKde*, **24**, 65 (1904).
(18) Steinhaus, W. *Phys. Z.* **12**, 657 (1911).
(19) Clayeux. *Onde élect.* **3**, 513 (1924).
(20) Jouaust, R. *C. R. Acad. Sci., Paris*, **175**, 1050 (1922).
(21) Pession, G. and Gorio, T. *Elettrotecnica*, **15**, 870 (1928).
(22) Hazel, H. C. *Proc. Inst. Radio Engrs*, **16**, 70 (1928).
(23) Martyn, D. F. *Proc. roy. Soc. Edinb.* **50**, 166 (1930).
(24) Barlow, H. E. M. *J. Instn elect. Engrs*, **77**, 612 (1935).
(25) Campbell, A. and Dye, D. W. *Proc. roy Soc.* A, **90**, 621 (1914).
(26) Albrandt, R. E. *Elektrotech. Z.* **56**, 340 (1935); *Izvestia Elektropromischleuosti Slaboga toka* (*Leningrad*). **7**, 24 (1934); **9**, 62 (1934).
(27) Matthias, A., Borries, B. von and Ruska, E. *Z. Phys.* **85**, 336 (1933).
(28) Brüche, E. *Z. Phys.* **88**, 205 (1934).

(29) Pierce, G. W. *Phys. Rev.* **19**, 196 (1904); **20**, 220 (1905).
(30) Keinath, G. *Jb. drahtl. Telegr.* **11**, 43 (1916).
(31) Turner, H. M. and Michel, P. C. *Proc. Inst. Radio Engrs*, **22**, 680 (1934).
(32) Moullin, E. B. *Proc. roy Soc.* A, **121**, 41 (1928); *J. Instn elect. Engrs*, **68**, 544 (1930).
(33) Fortescue, C. L. and Moxon, L. A. *J. sci. Instrum.* **8**, 94 (1931).
(34) Takaya, M. *Researches Electrotech. Lab. Tokyo.* No. 329 (1932).

§ 3. ELECTRICAL OSCILLATIONS AND BIOLOGICAL MATERIALS

By W. H. Ward

In recent years a good deal of work has been published on the effect of electric fields on biological material, especially at radio-frequencies. The interest of biologists in such work is obvious, and that so well-known a physicist as Prof. J. C. McLennan should have done some of it shows that the physics involved may be interesting. A biological material, in particular a part of an animal body, may be treated as a dielectric of very high power factor, and the results of applying to it an electric field are special cases of normal dielectric phenomena. The commonest application uses the dielectric loss of the human body in a strong electric field to heat the body. When the oscillations of the field have a frequency of a few million cycles per second the process is known as *diathermy*. When they have a frequency between ten and a hundred million cycles per second, it is known as *short-wave therapy*.

The published work is divisible into two parts; the one describing the use and applications of the clinical process, and the other describing observations and calculations intended to explain the physical processes behind the clinical one. Of the first part little need be said here. The standard accounts, referring particularly to short-wave therapy, are the books of Schliephake[1] and of Holzer and Weissenberg[2].

In the laboratory work, use is made of the similarity in dielectric properties between biological materials and dilute solutions of electrolytes to deduce from measurements on solutions the laws of heating of a biological material in an electric field. Since the result that is wanted is a quantitative value in heat units, it is unfortunate that very little seems yet to be known of the thermal properties involved. Most workers have applied an electric field to a material, and have made deductions from the resulting temperature rise, afterwards bridging the gap between the calculated dielectric loss and the observed change of temperature by arbitrary constants. The biological materials of the animal body have very high dielectric constants ϵ (up to 80), and conductivities σ of the order of thousandths of a mho per centimetre cube. From theoretical considerations confirmed by measurements on salt solutions, it has been established[3, 4, 5] that at a constant frequency f and constant current the loss in a material is a maximum if σ is made equal to $(\epsilon f/1{\cdot}8) \times 10^{-12}$ or $\tan \delta = 1$. For variations of σ at constant voltage, and for variations of frequency at either constant voltage or constant current there is no maximum of loss.

When two materials are in the same electric field it is possible for their relative

losses to vary with frequency. If they are in parallel the case is quite simple, for they may be considered as independent to a first approximation. If they are in series there is a frequency $f_1 = \omega_1/2\pi = \sqrt{\sigma_1\sigma_2}\,(\sigma_2 - \sigma_1)/(\sigma_2\epsilon_1^2 - \sigma_1\epsilon_2^2) \times 1{\cdot}8 \times 10^{12}$ at which the loss per unit volume in the two materials is equal. At lower frequencies the material of lower conductivity has the higher loss and vice versa. For frequencies less than f_1 the difference between the two losses increases as the frequency falls towards zero. For frequencies above f_1 there is a frequency f_m at which the difference between the two losses is a maximum; f_m is defined by

$$\frac{f_m^2}{f_1^2} = 1 + \frac{\epsilon_2}{\epsilon_1}\sqrt{\frac{\sigma_1}{\sigma_2}}\cdot\left(1 - \frac{\epsilon_1^2\sigma_2^2}{\epsilon_2^2\sigma_1^2}\right)\Big/\left(\frac{\sigma_2}{\sigma_1} - 1\right).$$

The relatively high loss in a material of low conductivity at low frequencies is well illustrated by the common statement[1, 3, 6, 7] that in short wave therapy (i.e. at higher frequencies) the body is heated throughout, whereas in diathermy the heating is more pronounced in the subcutaneous fat, which has a low conductivity. An approximately cylindrical volume of material of high dielectric constant surrounded by air distorts an electric field to such an extent that relative losses may at times be controlled by the shapes in which the materials are presented, and it is probable that this, too, may have considerable bearing on the observed differences of the distribution of heat at high and low frequencies.

It has so far been assumed that the dielectric constant and conductivity of a single biological material do not change with frequency. Actually, from its very nature the material is heterogeneous, so that, following Maxwell and Wagner, the effective dielectric properties vary with frequency, and one author[8] has claimed that measurements made on blood are well expressed by the equations

$$\epsilon = \epsilon_0\,[1 + \kappa/(1 + \omega^2\tau^2)], \quad \text{and} \quad \sigma = \sigma_0 + (\epsilon_0\kappa/4\pi\tau)\,[\omega^2\tau^2/(1 + \omega^2\tau^2)].$$

These two equations, well known in dielectric theory, refer to a suspension of conducting spheres in a non-conducting plasma. More comprehensive equations have been deduced[9, 10] and have also been found to fit observations. It is probable that similar equations[11, 12] will be the inevitable result of the cellular structure of all biological materials. The properties at high frequencies will then be determined by direct measurement more easily than by calculation from low-frequency values.

A suspension of bacteria is to the biologist a material of special interest and several authors have worked on its high-frequency characteristics. Unfortunately their results are still inconclusive. Krasny-Ergen[13] and Malow[14] have investigated, theoretically and experimentally, the possibility of making the individual bacteria hotter than their surrounding liquid. Their conclusions disagree. It has been reported that in a strong electric field at high frequencies multiple forms of bacteria break into simpler forms[15], and that toxins may be attenuated[16]. These phenomena, if they survive the criticisms[17] they have provoked, could readily be explained if the bacteria were hotter than the surrounding liquid.

Since the dielectric constant of biological materials in general is high, it is natural that there should be found in dealing with them marked distortions of the field. Edge effects, fringing, and electrode effects are very well known in measure-

ments of dielectrics. Air, biological material, and a metal have dielectric constants of about 1, 80, and infinity. The refraction at each boundary, and the wide change of reactance in crossing it cause a concentration of current at points and in small bodies of higher dielectric constant. Such concentrations are commonly used to produce local heating, by shaping the electrodes[18], and in a variety of other ways[19]. In short wave therapy it is the practice to separate the electrodes from the patient by a few centimetres of air, so keeping away from the skin the concentration of current near the edge of the metal. The air condenser so formed between the electrode and the patient has a reactance that is much greater than that of the biological material, and the current flowing through the body depends little on its conductance. As a result, for constant voltage between the electrodes, as well as for constant current, the loss in an organ tends to be greatest if $\tan \delta = 1$. The air gap also reduces the loss, otherwise serious, in drops of sweat on the skin.

A difficulty arising from distortion of the field is found in measuring the temperature of a material in an electric field. The glass of a thermometer has some dielectric loss, but it has a much lower dielectric constant than the materials in which it is immersed, while mercury, and the metal of a thermocouple, have very high dielectric constants. In either case, then, the field near the instrument is distorted, and the temperature rise recorded is not that of the material in general. The suggestion, made by several authors, of using a benzine-in-quartz thermometer is hardly an improvement. It certainly does not introduce appreciable loss, but its low dielectric constant must cause a distortion of the field, and the temperature is probably underestimated. If it is necessary to use a thermometer in the field, it should have dielectric properties closely resembling those of the test material. Alternatively heat sensitive paints[13], or the known reactions to temperature of various organic materials[14], may be used.

Numerous cases are reported[1, 20, 21, 22, 23] of increases of the rate of metabolism of various organisms in a strong electric field at high frequencies. In some cases[17, 24, 25, 26] it has been shown that exactly the same effect, but usually not so quickly, can be produced by normal heating. The absence of data on the thermal properties of the organisms, and particularly on the cooling of a part of the animal body by the blood stream, prevents a complete solution of several problems. Nerves have particularly interesting reactions to the electric field. Their material is said[15, 19] to heat very quickly, and the artificial fever which results on applying a strong electric field to the whole body[27] (the human body may be raised to a temperature of 105° F. in an hour) is likely to be useful in elucidating the mechanism of temperature-control[1, 19, 23, 26] of the animal body. There are other results[4, 22] with nerves that may not be due to heat. There is a certain amount of evidence, not conclusive, that an electric field may have a stimulating or a sedative effect on the nervous system, and it is reported[28] that a very weak field, wave-length 11 metres, caused fatigue of frog nerve-muscle preparation, and that the same effect may be observed at 300 m.[29], and at the positive pole with d.c. It is said that these changes correspond to changes in the colloidal condition of the sarcoplasm. These various effects on nerves may perhaps be due to a rectified alternating wave inter-

fering with the normal electrical activity of the nerve. The structure of a nerve, in which a semiconductor is enclosed in a sheath of lower conductivity, seems favourable to rectification. There may be some analogy with an observation on plants[30]. When a high-frequency magnetic field was around a plant, the sap rose very rapidly, the top of the plant became negative and the root positive, and the consequent current along the stem in the end killed the plant.

References

(1) Schliephake, E. *Short Wave Therapy* (1935). London: Actinic Press, Ltd.
(2) Holzer and Weissenberg. *Foundations of Short Wave Therapy* (1935). London: Hutchinson & Co.
(3) McLennan and Burton. *Canad. J. Res.* **3**, 224 (1930); **5**, 550 (1931).
(4) Hosmer. *Science*, **68**, 325 (1928).
(5) Pätzold. *Z. f. Hochfrequenztech.* **36**, 85 (1930).
(6) Réchou, Wangmerez, Halphen, Auclair and Dausset. *J. Radiol. Électrol.* **17**, 481 (1933).
(7) Malow. *Phys. Z.* **34**, 880 (1933).
(8) Rajewsky and Schaefer. *Kolloidzschr.* **77**, 257 (1936).
(9) Dänzer. *Ann. Phys., Lpz.*, **20**, 463 (1934); **21**, 783 (1934–5).
(10) Graul. *Ann. Phys., Lpz.*, **24**, 326 (1935).
(11) Malow. *Strahlentherapie*, **53**, 326 (1935).
(12) Cole. *Trans. Faraday Soc.* **33**, 966 (1937).
(13) Krasny-Ergen. *Z. Hochfrequenztech.* **48**, 126 (1936).
(14) Malow. *Phys. Z. Sowjet.* **9**, 271 (1936).
(15) McKinley. *Univ. Pittsburg Bull.* **30**, 183 (1933).
(16) Mellon, Szymanowski and Hicks. *Science*, **72**, 174 (1930).
(17) Curtis, Dickens and Evans. *Nature, Lond.*, **138**, 63 (1936).
(18) Nagelschmidt. *Brit. J. Radiol.* **8**, 449 (1935).
(19) Delario. *Radiology*, **25**, 615 (1935).
(20) Jellinek. *Brit. J. phys. Med.* **11**, 36 (1936).
(21) Denier. *J. Radiol.* **20**, 193 (1936).
(22) Schereschewsky. *Radiology*, **20**, 246 (1933).
(23) Ancelme. *Onde Elect.* **10**, 221 (1931).
(24) Hill and Taylor. *Lancet*, p. 311 (1936).
(25) Taylor. *Brit. J. Radiol.* **8**, 718 (1935).
(26) Latkowski and Charłampowicz. *Bull. int. Acad. Cracovie*, **4**, B, 11, p. 189.
(27) Carpenter and Page. *Science*, **71**, 450 (1930).
(28) Dalton. *Brit. J. phys. Med.* **11**, 221 (1933).
(29) Delherm and Fischgold. *J. Radiol.* **19**, 710 (1935).
(30) Marinesco. *C.R. Acad. Sci., Paris*, **194**, 1387 (1932).

§ 4. FRICTIONAL ELECTRICITY

By W. H. Ward

Frictional electricity was known in the remote past and fifty years ago it was shelved as too erratic for quantitative investigation. Recent work, however, shows some promise of an adequate explanation. High voltages, ascribed to friction, may be due to dustclouds[1], produced by a volcano or a tropical storm, to snow or hoar frost blown by the wind, to the flow of liquids[2] or gases[3] in pipes, to the movement[4] of photograph film in a camera, to the grinding of iron[5] on an emery wheel, to driving belts in dry situations, and to many other cases[6] of relative movement between two bodies. The voltage generated is usually harmful, and in

some trades Home Office regulations are framed to eliminate it. But the need for high voltages for experiments in atomic physics has revived the antique frictional generator. In one form a suspension[1] of dust[7] or liquid drops in air is blown past an electrode which is to be charged. In another a large paper[8,9,10,11,12] or rubber belt carries a charge from one pulley to apparatus at the other. The charge is produced either by friction at the first pulley or by an auxiliary voltage there. The second or insulated pulley may in favourable circumstances be raised to a potential of a million or so volts. This machine is a particularly clear case of the double process in the production of high frictional voltages. First an insulated body is charged by a relatively low voltage, and then it is carried to a conductor whose voltage is slowly built up. If the charged body is not well insulated in its travel from the charging point to the discharging point, its charge leaks away, and the voltage of the conductor is not increased. Instances of this are very common[10], beginning with the old frictional machine which worked badly on damp days—i.e. when various points of its glass plate had a conducting film of adsorbed water between them. It should be noted that the carriers themselves need not be good insulators—metal dust in air[1] is quite effective—but they must be insulated[3] from each other and from the charging point[13]. The potential of the conductor is actually the potential difference across the capacitance between the conductor and earth. If this capacitance is small the energy stored by it is unlikely to be destructive.

The rate at which the voltage of the conductor may be raised depends on the carrying capacity of the stream of carriers. For each of them, depending on its material, shape and size, there is a charge q which makes the potential gradient at the surface equal to the breakdown voltage[1,11,14,15] of the surrounding medium, usually air. The carrier may then have any charge up to q, but not more, and the maximum possible rate of increase of voltage of the capacitance C is then qnu/C, where n is the volume density of carriers, and u is the rate of flow of the stream. The maximum voltage attainable by the capacitance is limited either by the leakage from it, or by its own maximum charge Q for breakdown in its own medium.

When a dust cloud which has at some time moved in turbulent flow is analysed, it is found that its particles are usually charged[16], but that the numbers of particles of each sign are so nearly equal[17,18] that the resultant charge is often negligible. The charge on a particle is roughly proportional to its area and to its dielectric[15] constant. When the cloud is of liquid particles it is possible for the force within a drop due to its own electric charge either to control the size of the drop[19], or to cause it to burst[20].

The mechanism by which the insulating bodies which act as carriers become charged is not yet known with complete certainty. It is accepted that the origin of the charge lies in the contact[21] between two bodies, and that actual friction has no bearing whatever. The two explanations which have been advanced appear to be complementary parts of a complete explanation. The older one, of Helmholtz[1,19,22], is that when two dissimilar surfaces are in contact there exists between them an electrical double layer about 10^{-8} cm. thick. When the two surfaces are separated

the layer splits and a charge of one sign remains on each surface. In the course of the separation the distances between the two surfaces increases in a very large ratio, and the capacitance between them decreases correspondingly. Since the charge is constant, the potential difference between the two surfaces also increases by a very large ratio. The thousands of volts found on rubbing an insulator may be produced by a very low voltage across a very thin layer, multiplied as the rubber is removed. This theory is more often advanced to explain the charging of dielectrics. The second explanation is that the charging of two surfaces in contact is due to the contact potential difference[21,23,24,25], identical with the difference between the work functions[26,27] of the two materials. A connexion has been suggested between the generation of frictional electricity by a metal and the number of electrons[28] in the outer orbits of its atom. The results obtained in work on contact potential difference suggest several causes for the very erratic behaviour of frictional electricity. A metal surface, unless it has been heated strongly in a vacuum of the order of 10^{-8} mm. of mercury, gives results[26,27,29,30] that are due to adsorbed gases rather than to the metal. A monomolecular film of a fatty acid[31] on a surface has also a large influence on the contact potential, apparently connected with the orientation[32] of the dipoles of the film. Other very large changes are found to depend on the state of the surfaces in contact[26,33], and particularly on the distortion due to even the slightest cold working. Polished[24,25] and etched surfaces are not the same, nor are two different faces of a crystal[34,35]. There is some evidence of a connexion between contact potential and dielectric constant[19,21,22].

References

(1) VOLLRATH. *Phys. Rev.* **42**, 298 (1932).
(2) REICHART. *C.R. Acad. Sci., Paris*, **202**, 1494 (1936).
(3) RAYNER. *Lancet*, part 2, 1408 (1936).
(4) ROSE. *Canad. J. Res.* **9**, 197 (1933).
(5) NAKAYA. *Bull. Inst. phys. chem. Res. Japan*, **23**, 185 (1934).
(6) *Industrial Chemist*, pp. 80, 104 (1936).
(7) BURKHARDT, *Ann. Phys., Lpz.*, **23**, 339 (1935).
(8) HERB, PARKINSON and KERST. *Phys. Rev.* **51**, 75 (1937).
(9) BARTON, MUELLER and VAN ATTA. *Phys. Rev.* **42**, 901 A (1932).
(10) VAN ATTA, NORTHRUP, VAN ATTA and VAN DER GRAAFF. *Phys. Rev.* **49**, 761 (1936).
(11) VAN DER GRAAFF, COMPTON, and VAN ATTA. *Phys. Rev.* **43**, 149 (1933).
(12) BRAMHALL. *Rev. sci. Instrum.* **5**, 18 (1934).
(13) BÜHL. *Kolloidzschr.* **59**, 346 (1932).
(14) PAUTHENIER and MOREAU-HANOT. *C.R. Acad. Sci., Paris*, **201**, 1332 (1935).
(15) CHAPMAN. *Physics*, **5**, 150 (1934).
(16) NAKAYA and TERADA. *Phil. Mag.* **19**, 115 (1935).
(17) SACHSSE. *Ann. Phys., Lpz.*, **14**, 396 (1932).
(18) MILHOUD. *C.R. Acad. Sci., Paris*, **198**, 1586 (1934).
(19) LEWIS. *Trans. Faraday Soc.* **28**, 597 (1932).
(20) ZELENY. *J. Franklin Inst.* **219**, 659 (1935).
(21) RICHARDS. *Phys. Rev.* **22**, 122 (1923).
(22) HESS. *Z. Phys.* **78**, 7–8, 430 (1932).
(23) MAINSTONE. *Phil. Mag.* **19**, 278 (1935).
(24) MAINSTONE. *Phil. Mag.* **12**, 535 (1931).
(25) HANSTOCK. *Phil. Mag.* **13**, 81 (1932).

(26) ANDERSON. *Phys. Rev.* **47**, 958 (1935).
(27) GLASOE. *Phys. Rev.* **38**, 1490 (1931)
(28) SHAW and LEAVEY. *Proc. roy. Soc.* A, **138**, 502 (1932).
(29) LINTERN and ADAM. *Trans. Faraday Soc.* **31**, 564 (1935).
(30) REIMANN. *Phil. Mag.* **20**, 594 (1935).
(31) SCHULMAN and HUGHES. *Proc. roy. Soc.* A, **138**, 430 (1932).
(32) ADAM and HARDING. *Proc. roy. Soc.* A, **138**, 411 (1932).
(33) BRADFORD. *J. chem. Soc.* p. 1276 (1934).
(34) ROSE. *Phys. Rev.* **44**, 585 (1933).
(35) ZISMAN and YAMINS. *Phys.* **4**, 7 (1933).

ELECTROLYTES AND ELECTROLYSIS

By R. W. GURNEY, Ph.D.

§ 1. INTRODUCTION

Although Faraday put forward the idea that during electrolysis the current is carried by charged particles or ions, he did not insist or suggest that these positive and negative ions continued to exist separately in the solution after the applied potential difference had been removed. For many years it was supposed that when a potential difference was applied to an electrolytic solution, a certain proportion of the solute molecules were broken down by the applied field, and that the ions recombined with one another when the field was cut off, although during the passage of the current the ions moved independently of one another. On this view strongly conducting solutions were given by those substances of which a relatively large amount was temporarily broken down into ions by the applied potential. If this view were correct it would be expected that a definite minimum potential difference would be necessary to decompose the solute molecules; but it was found that in many cases a measurable current flowed, no matter how small was the applied potential difference. This fact, among others, led Arrhenius in 1883 to suggest that in any conducting solution a certain number of electrically charged ions exist permanently, a definite fraction of the solute having become dissociated into ions at the moment when it was dissolved in the solvent. When a small potential difference is applied to such a solution, the positive and negative ions drift in opposite directions, no additional molecules being broken down into ions by the field. Weakly conducting solutions are given by those substances which exist in solution mainly as neutral molecules, only a small number of molecules having been broken down into ions.

At about the same time Van't Hoff had discovered that the ordinary gas laws were often applicable to solutes in dilute solution; for many dissolved substances the osmotic pressure was found to have the same value as the gas pressure which the substance would exert if it were a gas at the same temperature and volume. Some substances, on the other hand, were found to exert osmotic pressures greater than the expected value. Arrhenius, noticing that these anomalous solutions were all electrically conducting solutions, recognized that these high osmotic pressures were to be expected if his idea of permanent dissociation were correct; for each positive and each negative ion would contribute to the pressure as much as a single molecule. (It should be understood that these osmotic pressures are for the most part not measured directly but are deduced from such observations as the lowering of the freezing-point, etc.) Solutions of $BaCl_2$, for example, were found to have osmotic pressures more than 2·5 times the normal value; which could be explained if some of the $BaCl_2$ molecules were dissociated into as many as three ions. For a solute like KCl, on the other hand, which will dissociate into not more than two ions, the

osmotic pressure should never be more than twice the value for a normal undissociated solute. The experimental values were found to be consistent with these rules. For all electrolytes dissolved in alcohol or similar solvent the osmotic pressure was found to be nearer the normal value, indicating a lower degree of dissociation than in the corresponding aqueous solution.

At the time when Arrhenius proposed these ideas of permanent ionic dissociation, it had already been recognized for some twenty years that the state of a partially dissociated gas depended upon the volume in which it was contained, the degree of dissociation α being the smaller, the higher the density, the variation being given by

$$\frac{\alpha^2}{1-\alpha} = Av \qquad \text{.....(1),}$$

where v is the volume and A is a constant. To any partially dissociated solute obeying the gas laws the same Law of Mass Action must apply, the degree of dissociation into ions being the smaller, the higher the concentration of the solute. When the lowering of the freezing-point of aqueous solutions of electrolytes was examined, the results indicated that the expected type of behaviour was present in every case; see table 1. Measurements of the vapour pressure of the solutions gave the same results within the experimental error.

Since both these methods depended on a single property of the solution—its osmotic pressure—it was desirable to have an independent test of the degree of ionic dissociation from a study of some property having no connexion with osmotic pressure. The electrical conductivity of dilute solutions, depending as it must on the number of ions present, appeared to provide another direct test of the degree of ionic dissociation at different concentrations. In a very dilute aqueous solution each solute particle moves among water molecules, and only rarely encounters another solute particle. Consequently in a sufficiently dilute solution an ion is moving among water molecules, and the viscous forces limiting its mobility will be almost identical with those for the ion in pure solvent. Ostwald supposed, therefore, that in very dilute solutions the mobility could be taken as independent of the concentration; in which case the values of the conductivity would be a direct measure of the number of free ions present, and hence of the degree of ionic dissociation of the solute. The equivalent conductivity Λ should increase towards a limiting value Λ_∞ at infinite dilution, corresponding to the condition where all the solute molecules are dissociated into ions. From equation (1) Ostwald's expression for Λ follows at once:

$$\frac{\Lambda^2}{\Lambda_\infty(\Lambda_\infty - \Lambda)v} = \text{constant} \qquad \text{.....(2).}$$

For all feebly conducting solutions, at any rate, the behaviour of the conductivity was in satisfactory agreement with this law. For electrolytes dissolved in alcohol, etc., the conductivity was always found to be smaller than for the corresponding aqueous solution. This agreed with the values of the osmotic pressures, mentioned above, which indicated a lower degree of dissociation into ions.

As soon as a certain degree of permanent ionic dissociation was widely accepted, rapid progress was made in the interpretation of many aspects of electrolytic solutions. Nernst, for example, soon used the idea in his well-known mechanism for the electromotive force of cells. When a metal, such as silver, is dipped into a concentrated solution containing its own ions, a few of the latter may be deposited on the metal, charging it positively—or rather, setting up an electrical double layer on the surface, negative outwards and positive inwards. The strength of this double layer being determined by the number of ions in solution, Nernst was led to propose his law for the dependence of e.m.f. on concentration. During the same period similar advances were made in the understanding of such important phenomena as the ionic dissociation of pure water, the behaviour of oxidation-reduction cells, etc., with the result that before the end of the nineteenth century electrochemistry had already been given that framework within which investigation has been carried on until the present day.

§ 2. IONIC STRUCTURE

This theory of Arrhenius was put forward many years before the electron had been isolated as an elementary particle capable of independent existence. With the discovery of the electron people learned to recognize a positive ion as an atom or molecule from which one or more electrons have been removed, and a negative ion as a particle to which one or more additional electrons have become attached; and in the neutralization of an ion at an electrode, when we speak of a positive ion "giving up its charge to the metal" we recognize that what we really mean is that the ion captures one or more electrons from the metal.

With the introduction of the conception of the nuclear atom these ideas became more precise. The charge on a positive atomic ion is due to the fact that the number of orbital electrons present is for the time being insufficient to neutralize the positive charge on the nucleus; and the charge on a molecular positive ion is due to the fact that the total number of orbital electrons is insufficient to neutralize the sum of the charges on the atomic nuclei in the molecule.

When to each element in the Periodic Table was assigned its Atomic Number, definite relationships between various common species of ions were recognized. Thus the ions Mg^{++}, Ca^{++}, Sr^{++}, and Ba^{++} have the same number of extra-nuclear electrons, and the same electronic structure as the alkali ions Na^{+}, K^{+}, Rb^{+}, and Cs^{+}, respectively; both, in fact, have the closed electronic shells of the rare gas atoms Ne, Ar, Kr and Xe, respectively. Since the same is true of the halogen negative ions F^{-}, Cl^{-}, Br^{-}, and I^{-}, and of such ions as Al^{+++}, it appears that many of the familiar ions in aqueous solution have the completed electronic shells of the rare gas structure.

The introduction of the Bohr model for atoms and ions had, however, few repercussions on the study of ions in solution. The absorption spectra of solutions not being amenable to any simple interpretations, attention continued to be focused upon the macroscopic properties of solutions. Although the energies of ions in gases have long been expressed in electron-volts per particle, the energies of ions in

solution are still usually given in calories per mol, which typifies the different approach. Within the framework which had been built up by the end of the nineteenth century more systematic and more accurate measurements of conductivities, of e.m.fs., and activity coefficients, were made; and it became increasingly clear that, though for feebly conducting solutions the theory was satisfactory, for all strong electrolytes the so-called equilibrium between ions and undissociated molecules showed no quantitative agreement with the laws of Arrhenius and Ostwald at all, though their behaviour was always qualitatively in the right direction.

After 1919 the view became prevalent that in the solid state many of the common substances, like NaCl, consist entirely of positive and negative ions arranged in a lattice; and this knowledge led people to consider once more the possibility that in dilute solution such substances consist entirely of positive and negative ions—a suggestion that had already been made several times. If one adopts this point of view—that in dilute solutions the solute is completely dissociated into ions at all concentrations—one has of course to find a new explanation for the variation of the osmotic pressure and of the electrical conductivity with change of concentration. For solutes which give feebly conducting aqueous solutions the original viewpoint of Arrhenius has never been questioned—in dilute solution these substances consist mainly of neutral molecules. For strong electrolytes, on the other hand, it is now recognized that the effects which were previously attributed to a varying proportion of undissociated molecules are due instead to the electrostatic forces between the ions, there being in dilute solution no neutral molecules at all.

§ 3. INTERIONIC FORCES

Even in very dilute solutions the paths followed by the ions in their Brownian movement will to some extent be dictated by the electrostatic forces. The repulsion between ions of the same sign will oppose their approach, while the attraction between ions of opposite sign will encourage their approach. The result is that, if we fix attention on any positive ion in solution and observe its nearest ionic neighbours over a certain period of time, we shall find among them slightly more negatives than positives; while if we fix attention on any negative ion, we shall find among its neighbours more positives than negatives. (We have here a very slight tendency towards the kind of ionic arrangement which occurs in a rock-salt crystal, where each positive ion is surrounded by six negatives, and each negative is surrounded by six positives.) Consider then any ion in dilute solution; to remove this ion from among its neighbours a certain amount of work will have to be done, since among the neighbours is an excess of ions of the opposite sign.

Now there is a well-known theorem in electrostatics, that if we have any group of point charges in fixed positions, and if the work to remove the first charge to infinity is w_1, and the work to remove the second charge when all the other charges, including the first, are present, is w_2, and so on, then the work to separate the whole system of charges and leave them at infinite distances from one another is

$$\tfrac{1}{2}(w_1+w_2+w_3+\ldots)=\tfrac{1}{2}\sum_i w_i. \qquad \ldots\ldots(3).$$

The existence of this electrostatic energy will clearly affect the osmotic pressure of any electrostatic solution. For if the solute is contained in a large volume v, and we diminish this volume by using a semi-permeable membrane, the solution will fall to a state of lower electrostatic energy. Consequently, instead of having to do the normal amount of work

$$P_0 dv = \frac{RT}{v} dv = \frac{nkT}{v} dv \qquad \text{.....(4)}$$

we shall have to do a smaller amount

$$P_0 dv = P_0 dv - dF_e \qquad \text{.....(5),}$$

where dF_e is the change in the electrostatic contribution to the free energy.

The original interest in table 1 lay in the large factors by which the values of i in each column exceeded unity. The interest now has shifted; we wish to explain why the figures for KCl and $MgSO_4$ are less than 2; and why the figures for $BaCl_2$ and $La(NO_3)_3$ are less than 3 and 4 respectively. We attribute this to the presence of the electrostatic term F_e, which we proceed to evaluate from the interionic forces.

Table 1. γ, concentration (mol per 1000 g. of water); i, ratio of observed osmotic pressure to the value for undissociated molecules

KCl		$MgSO_4$	
γ	i	γ	i
0·0010	1·97	0·0049	1·60
0·0100	1·94	0·0130	1·48
0·0867	1·86	0·0278	1·39
$BaCl_2$		$La(NO_3)_3$	
γ	i	γ	i
0·00087	2·86	0·00132	3·75
0·0114	2·72	0·00354	3·56
0·1136	2·52	0·00806	3·40

Electrolytic solutions may be divided into two classes, those where the positive and the negative ions are equal in number, as in a solution of $MgSO_4$ or KCl, for example, and those where the numbers are unequal, as in any uni-divalent solute such as K_2SO_4. Most of the properties of solutions have been worked out for both kinds, but the physical principles always appear most clearly in discussions of the first simpler class. In this chapter we shall restrict our attention to binary electrolytes dissociating into ions bearing equal charges $\pm q$, where $q = z\epsilon$, ϵ being the electronic charge and z an integer. The corresponding expressions for the other class of electrolytes will be found in the journals referred to.

In any liquid at temperature T the average thermal energy of a molecule has a definite value; and in a dilute solution each solute particle is in thermal equilibrium with the adjacent solvent molecules. The solvated ions will be continually moving about with Brownian motion, their paths being determined by the equally random

motion of the surrounding molecules. In this motion each positive ion will from time to time approach a negative ion, and will afterwards leave it again. Now in a medium of dielectric constant K the force of attraction between unlike charges $\pm q$ at a distance d apart is q^2/Kd^2; and to separate this pair of charges against their mutual attraction an amount of work q^2/Kd must be done. When in their Brownian motion a positive and a negative ion approach one another they lose electrostatic potential energy, and the kinetic energy that they acquire is communicated to the solvent molecules. Conversely, when in their Brownian motion a pair of opposite charged ions move apart against their own mutual attraction, they receive the necessary energy from the thermal energy of the adjacent solvent molecules; and we are led to enquire how large in typical solutions will be the value of q^2/Kd compared with the thermal energy of a molecule in a liquid at room temperature, which we are accustomed to measure in units of RT per mol, or kT per particle. For when the distance between a pair of ions is large, the energy necessary to separate them will be small compared with kT, and these ions can move apart at any moment without difficulty. In a very dilute solution there will be few ions for which this condition is not realized, and for this reason solutions at very low concentration present a simpler problem than solutions less dilute; for solutions where K is small and q is large the concentration will need to be considerably lower than for univalent electrolytes in water, where K is large and q is small.

In 1923 Debye and Hückel[1] worked out the problem of the interionic forces under the assumption that q^2/Kd is small compared with kT for all ions in the solution, and three years later, in 1926, Bjerrum[2] described the state of a solution where this condition is not satisfied. Bjerrum's description, which is specially important in solvents of low dielectric constant, was later confirmed by the detailed statistical treatment of Fuoss[3].

§ 4. THE DEBYE-HÜCKEL THEORY

Suppose that we fix attention on a small volume dv at a fixed point in the interior of a dilute solution. Sometimes a positive ion and sometimes a negative ion will enter this volume; but we know that if our observations extend over a period of time long compared with that of the Brownian motion, the average charge density will be zero. Suppose, on the other hand, that we observe a small volume dv, not at a fixed point, but at a short distance r from one particular ion, on which we choose to fix attention; we shall call this ion the "central" ion; as this ion moves about the volume dv moves with it at a fixed distance r. We shall show that under these circumstances the average charge density in the volume dv is not zero, but is negative when the central ion is positive, and is positive when the central ion is negative. At large distances the central ion will exert a negligible influence; hence with increasing r the charge density ρ thins out, and tends to zero.

If V_+ is the potential energy of a positive ion in the volume dv, the average number of positive ions in this volume will be, by Boltzmann's law,

$$n_+ dv = ne^{-V_+/kT}\, dv \qquad \text{.....(6).}$$

Similarly, if V_- is the potential energy of a negative ion in the same volume element, the average number of negative ions will be

$$n_- dv = ne^{-V_-/kT}\, dv \qquad \text{.....(7).}$$

If the temperature is raised, the distribution approaches a uniform random distribution; both n_+ and n_- tend towards n, the number of ion pairs per unit volume.

The average charge density is given by

$$\rho = (n_+ q - n_- q) \qquad \text{.....(8),}$$

the sign of ρ being positive or negative according as equation (6) is greater or less than equation (7). As we are dealing with purely electrostatic forces, $V_+ = q\psi$ and $V_- = -q\psi$, where ψ is the potential of the electrostatic field. This ψ consists of two parts, the part $q/K\epsilon$ due to the central ion, and a smaller part due to the neighbours,

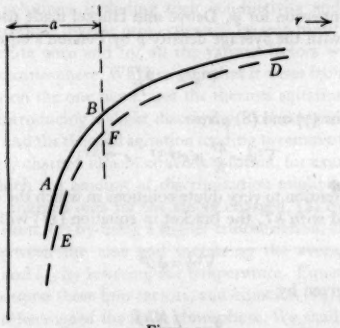


Figure 1

which we are trying to evaluate. The nature of the modification can perhaps most easily be seen by asking what would be the potential energy of an electron passing through this field. Let the broken curve *EFD* of figure 1 be a plot of $-\epsilon q/K\epsilon$, the potential energy which the charge ϵ would have in the field of the central ion if the other ions were absent; and consider how the modified curve will differ from this. For large values of ϵ the curve must still become asymptotic to the axis; and on the other hand, when the electron is extremely near the central ion it is in a field which has the same intensity as if the neighbours were absent. For the central solvated ion has a certain radius, and no neighbour on collision will obtrude its charge within a sphere that has a certain radius a. For all values of ϵ less than a the required curve must have exactly the same slope as if the neighbours were absent; i.e. in this region the curve is the same as *EF*, only shifted vertically upward or downward. Clearly the shift must be an upward one; for the slightly negative atmosphere round the central positive ion exerts a screening effect; the potential energy curve will therefore fall less steeply than *DF*. We require some curve like *DBA*, whose shape will depend upon the concentration of the solution. We shall have to obtain an expression

for the curve DB; but our only object in doing so will be to evaluate the vertical distance of AB above EF. For if we consider what this vertical shift represents, we shall see that it gives us the quantity that we require. The curve EF was a plot of $-\epsilon q/Kr$, the potential energy of our electron in the field of the charge q alone. The curve AB gives the total potential energy-$e\psi$ in the field of the central charge and its neighbours. If the vertical shift of AB above EF is called $\epsilon\chi$

$$\epsilon\psi = \epsilon\,(q/Kr - \chi) \qquad \text{.....(9).}$$

The quantity χ is the constant electrical potential within this region due to the neighbours outside. Since the central ion itself is not subject to its own field, but only to that of its neighbours, its potential energy is just $-q\chi$. This is the quantity that we are looking for; the work required to remove any ion from among its neighbours.

To obtain an expression for ψ, Debye and Hückel made use of the fact that it must be connected with the average density ρ by Poisson's equation

$$\nabla^2\psi = -\frac{4\pi\rho}{K} \qquad \text{.....(10),}$$

which from equations (7) and (8) gives

$$\nabla^2\psi = \frac{4\pi}{K}\,n(e^{q\psi/kT} - e^{-q\psi/kT}) \qquad \text{.....(11).}$$

If we now restrict attention to very dilute solutions in which the electrostatic energy $q\psi$ is small compared with kT, the bracket in equation (11) will reduce to $2q\psi/kT$, and may be written

$$\nabla^2\psi = \psi/l^2 \qquad \text{.....(12),}$$

where l is a length given by

$$l = \sqrt{\frac{KkT}{8\pi nq^2}} \qquad \text{.....(13).}$$

The solution of equation (12) which tends to zero for large r is

$$\psi = \frac{Ae^{-r/l}}{r} \qquad \text{.....(14).}$$

When the value of the constant A has been correctly determined, we can obtain an expression for the curve BD in figure 1. Now the intensity of the electrostatic field in this region is found by differentiating equation (14) with respect to r, which gives

$$\frac{d\psi}{dr} = -\frac{Ae^{-r/l}}{r^2}\,(1 + r/l) \qquad \text{.....(15).}$$

In particular the intensity of the field at $r = a$ can be obtained by substituting in this expression. But we already know the intensity of the field here; in the region AB of figure 1 the field is simply q/r^2 due to the central charge, and is consequently q/a^2 at $r = a$. Equating these two expressions we find

$$A = \frac{q}{K}\,\frac{e^{a/l}}{(1 + a/l)} \qquad \text{.....(16).}$$

Writing $r=a$ in equations (14) and (15) and substituting for A, we find

$$qX=q^2\left(\frac{1}{Ka}-\frac{1}{Ka\,(1+a/l)}\right)=\frac{q^2}{Kl\,(1+a/l)} \qquad \text{.....(17).}$$

In the very dilute solutions with which we are dealing the radius a is small compared with the length l, and a/l may be omitted from the denominator. We find then that the work required to remove an ion from its neighbours when there are n ion pairs per cm.^{3} is simply

$$qX=\frac{q^2}{Kl}=\frac{q^3\,n^{\frac{1}{2}}}{K^{\frac{1}{2}}}\sqrt{\frac{8\pi}{kT}} \qquad \text{.....(18).}$$

We could now use this result to discuss the osmotic pressure of dilute solutions. But in subsequent sections the ionic atmosphere will be used to interpret several other properties of solutions, including their conductivity and viscosity; and the relationships between these diverse properties will be seen more clearly if we stop at this point to enumerate once and for all the various factors which determine the structure of the ionic atmosphere. We have seen that it arises from a balance between the interionic forces on the one hand, and the thermal agitation on the other—the electrostatic forces introducing a slight discrimination between ionic neighbours of like and unlike sign, and the thermal agitation tending to remove this degree of order. If we start with singly charged ions in aqueous solution, for example, there are four possible ways in which the amount of discrimination might be increased: (1) by substituting doubly charged ions for the singly charged, (2) by using a solvent of lower dielectric constant, (3) by using a higher concentration, thus diminishing the average distance between the ions and increasing the average intensity of the electrostatic forces, and (4) by lowering the temperature. Equation (13) shows how the length l depends upon these four factors, and equation (14) shows how l defines the compactness or otherwise of the ionic atmosphere. We shall see in later sections how all the properties of solutions due to the presence of this ionic atmosphere depend upon the value of l, and consequently are governed by the four factors enumerated above. Using equation (3) we see that the electrostatic energy of the solution is proportional to q^3, to $K^{-\frac{1}{2}}$, and to $n^{\frac{1}{2}}$; and it can easily be shown that the value of the free energy is obtained by multiplying by $\frac{2}{3}$.

§ 5. COMPARISON WITH EXPERIMENT

We are now in a position to inquire whether the osmotic pressures of simple electrolytes can be accounted for by the assumption that in dilute solution they are completely dissociated into ions. For P_0-P is equal to dF_e/dv, which, we find, varies as q^3, as $K^{-\frac{1}{2}}$, and as $n^{\frac{1}{2}}$. Now from equation (4) P_0 is proportional to n, consequently the relative diminution $(P_0-P)/P_0$ should be proportional to $\sqrt{n}$, and if plotted against $\sqrt{n}$ or $\sqrt{\gamma}$ should give a straight line, passing through the origin.* For a substance like $MgSO_4$, whose ions are doubly charged, the slope, proportional

* The ratio P/P_0 for any solution is known as the osmotic coefficient, and is often denoted by g; the ordinates of figures 2 and 3 are $(1-g)$.

to q^3, will be eight times as steep as for a substance like KCl. These calculated slopes are shown in figures 2 and 3, together with the experimental points. At the lowest concentrations at any rate one can account for the observed values without the assumption of any undissociated molecules.

Figure 3 shows some results for potassium iodide dissolved in cyclohexanol[4]. The value of the dielectric constant of this solvent at room temperature is only 15; consequently the theoretical slope, shown in the diagram, is 12·4 times the slope for the corresponding aqueous solution (which is also shown). The experimental points show that at the lowest concentrations the solute appears to be completely dissociated into ions, even in this solvent of low dielectric constant. Presumably no solute which is wholly or partially dissociated into ions will exert a smaller osmotic

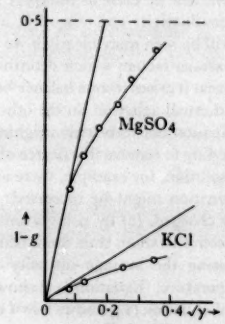


Figure 2

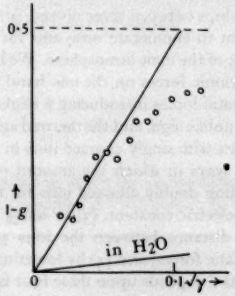


Figure 3

pressure than it would if it were not dissociated at all; if this is so, the value of P/P_0 for any electrolyte which furnishes a pair of ions cannot fall below $\frac{1}{2}$; that is to say, the value of $(P_0-P)/P_0$ will not rise above 0·5. In figures 2 and 3 a horizontal line has therefore been drawn at the value 0·5, and we must expect that the experimental curve, starting at the origin with the theoretical slope, will bend over so as to keep below this line. Solutions at these higher concentrations will be discussed below in §7.

§6. ELECTRICAL CONDUCTIVITY OF DILUTE SOLUTIONS

The Ostwald law for the conductivity of solutions seemed to show rather conclusively that in every electrolytic solution the number of ions varied with the concentration. Complete dissociation, however, means that the number of ions remains constant; and a totally different explanation for the conductivities of strong electrolytes had to be found. The partial discrimination between ionic neighbours of like and unlike sign was again found to provide the required mechanism.

From equations (10), (12) and (14) we see that at any distance r from the central ion the average charge density ρ has the value

$$\rho = \frac{qe^{-(r-a)/l}}{4\pi r l^2} \qquad \text{.....(19),}$$

the sign being the opposite to that of the charge of the central ion. With increasing r the charge thins out exponentially at a rate depending on the magnitude of l. For a uni-univalent solute in 1/1000-M. aqueous solution, for example, the value of l is about 10^{-6} cm.

Consider now what would happen if we could annihilate or suddenly neutralize the charge of the central ion at a particular moment. The ionic atmosphere due to this ion would have to disappear by a movement of the neighbouring ions. But since these ions take time to move, the atmosphere will take a certain time to disappear. At every point the value of ρ, given initially by equation (19), will fall everywhere towards zero at a rate which can be characterized by a time Θ which we call the "time of relaxation of the ionic atmosphere". The greater the electrostatic restraint which has been present, the more rapidly will the atmosphere be dispersed; Θ is, in fact, proportional to l^2, and depends also on the mobilities of the ions. For simple electrolytes the order of magnitude of Θ is found to be

$$\Theta \fallingdotseq \frac{10^{-10}}{\gamma} \text{ sec.} \qquad \text{.....(20),}$$

that is to say, about 10^{-7} sec. in 1/1000-M. solution.

Suppose now that a positive ion, initially at the point O in figure 4, has been moved rapidly to the point O^1. The ionic atmosphere centred round O will have begun to die away, but the negative charge density will not have had time to fall to its equilibrium value. Similarly, the negative charge density in front of the ion will not have had time to grow to its equilibrium value; that is to say, this region is more positive than it ought to be. If the positive ion moves along a line like OO', it therefore experiences a retarding force, due both to the attraction of the too negative region behind it, and to the repulsion from the too positive region in front. The greater the velocity of the ion, the greater will be this asymmetry; and investigation shows that the retarding field is proportional to the velocity v of the ion. Consider then the passage of a current through a solution to which a steady field E is applied. In this field the ion moves with a velocity $v = uE$, where u is the mobility; the retarding field is proportional to the applied field, and the current obeys Ohm's law. At great dilution the ionic atmosphere becomes less important, and the apparent mobility u tends to the limiting value u_∞. At any finite concentration the apparent mobility u differs from u_∞ because the ion tends to move backwards in the retarding field with the mobility u_∞. Thus we may write

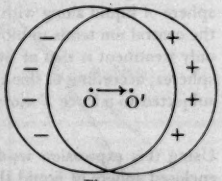


Figure 4

$$u_\infty - u = au_\infty \qquad \text{.....(21),}$$

where a is a constant, or in terms of conductivities

$$\Lambda_\infty - \Lambda = a\Lambda_\infty \qquad \text{.....(22).}$$

We see then that it was a mistake of Ostwald and Arrhenius to assume that in dilute solutions the equivalent conductivity is a direct measure of the number of ions present.

Now there is another mechanism by which the electrostatic forces affect the mobility of the ions. Let us again examine a solution through which a steady current is flowing, fixing attention on a particular positive ion which is drifting to the right in the applied field. Its neighbours of like sign will all be drifting in the same direction, while its neighbours of opposite sign will all be drifting to the left. Since these neighbours of opposite sign are in excess, and in their motion are to some extent carrying the solvent with them, the central ion is not situated in a stationary solvent, but in a solvent which is streaming all the time in the opposite direction. If we know the velocity of the ion relative to the immediate solvent, this does not tell us the rate at which the ion travels until we have calculated and subtracted the velocity of that solvent which is in immediate contact with the ion. This effect, which is known as the electrophoretic effect, further diminishes the conductivity to a degree depending upon the discrimination between like and unlike neighbours, that is to say, depending on the concentration of the solution.

To evaluate this effect the method used by Debye, Hückel and Onsager[5] is as follows. Divide the ionic atmosphere into concentric shells of thickness dr, and fix attention on one shell of internal radius r. The volume of this shell is $4\pi r^2 dr$, and the total charge contained in it is $4\pi\rho r^2 dr$. In an applied field E the electrostatic force acting on this shell is $4\pi\rho E r^2 dr$. This tends to move the shell and the enclosed sphere of liquid along with it, in precisely the opposite direction to that in which the central ion tends to move. For the motion of a sphere in a viscous medium the only treatment is that of Stokes, which was intended to apply only to macroscopic spheres; according to this treatment in a medium of viscosity η a sphere of radius r subjected to a force F moves with a velocity

$$F/6\pi\eta r \qquad \text{.....(23).}$$

Using this expression we find that each concentric sphere communicates to the enclosed sphere of liquid the velocity

$$dv = \frac{4\pi r^2 E\rho\, dr}{6\pi\eta r} = \frac{2}{3}\frac{E\rho r\, dr}{\eta} \qquad \text{.....(24).}$$

The total velocity in the wrong direction, communicated by the ionic atmosphere to the central ion, is obtained by integrating from $r = a$ to infinity

$$v = \int_{r=a}^{r=\infty} dv = -\frac{qE}{6\pi\eta l} \qquad \text{.....(25).}$$

This velocity with which the ion tends to move in the wrong direction is again proportional to the applied field E; the current obeys Ohm's law, and the effect may

be expressed as another negative conductivity, which may be written in the form $\beta\sqrt{\gamma}$, since the length l which occurs in equation (25) is inversely proportional to the square root of the concentration. Though the derivation is too long to be given here, it is found that the coefficient a in equation (22) is also proportional to $1/l$ and hence to $\sqrt{\gamma}$. The whole diminution in the conductivity from both causes may therefore be written in the form

$$\Lambda_\infty - \Lambda = a\Lambda_\infty\sqrt{\gamma} + \beta\sqrt{\gamma} \qquad \text{.....(26)}$$

$$= (\alpha\Lambda_\infty + \beta)\sqrt{\gamma} \qquad \text{.....(27).}$$

As long ago as 1900 an empirical law had been proposed by Kohlrausch that the diminution in the conductivity was proportional to $\sqrt{\gamma}$.

It should be noticed that the Debye-Onsager theory makes no attempt to say anything about the absolute values of the conductivities. But, given an empirical limiting value u_∞ for each species of ion, it states how the equivalent conductivity will fall with increasing concentration. The higher the initial value of Λ_∞ the more rapidly will the value fall as the concentration is increased; this is for the reason mentioned above, i.e. that in the asymmetrical field the ion tends to move backwards with its same characteristic mobility with which it moves forward in the applied field. On the other hand, the electrophoretic term contains nothing characteristic of the species of ion, except its charge. For uni-univalent electrolytes equation (27) takes the form

$$\Lambda_\infty - \Lambda = \left(\frac{5{\cdot}77\times10^5}{(KT)^{\frac{3}{2}}}\Lambda_\infty + \frac{58{\cdot}0}{(KT)^{\frac{1}{2}}\eta}\right)\sqrt{2\gamma} \qquad \text{.....(28).}$$

If the observed values of the conductivity are plotted against the square-root of the concentration, at low concentrations the points should lie on a straight line, whose slope is given by equation (28). For uni-univalent salts in water at 25° C., equation (28) takes the form $(0{\cdot}161\,\Lambda_\infty + 42{\cdot}28)\sqrt{2\gamma}$; in methyl alcohol ($K = 30$) the slope is much greater, namely $(0{\cdot}674\,\Lambda_\infty + 109)$, as shown in figure 5. In water the results for both uni-univalent and uni-divalent salts lie well on the theoretical lines. For di-divalent salts one must expect that lower concentrations would be necessary; earlier measurements for $MgSO_4$ showed a slope 70 per cent steeper than the theoretical; recent results of Sementschenko and Serpinsky[6] obtained at greater dilutions, however, have a slope only 30 per cent greater than the theoretical. Similarly, in solvents of low dielectric constant for many salts agreement with the theory is obtained only at very low concentrations. For silver nitrate in methyl alcohol[7], shown in figure 5, the deviation is exceptionally large. In all solvents the deviation of Λ from Λ_∞ becomes more rapid at higher temperatures; in equation (28)

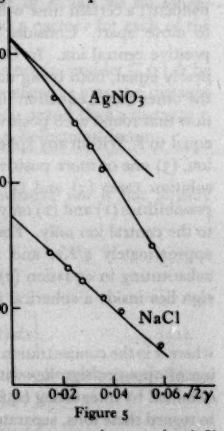


Figure 5

both the terms are increased, the former because of the rise in Λ_∞, and the latter because of the rapid fall in the value of η. The experimental results for uni-univalent salts are in satisfactory agreement with the theory.

§ 7. LESS DILUTE SOLUTIONS

Although in later sections we shall make other applications of the ionic atmosphere, it will be more convenient at this point to discuss less dilute solutions. As we saw above, Debye's treatment was restricted to solutions where the condition $q^2/Kd \ll 2kT$ is fulfilled for all but a negligible number of the ions present. Three years later Bjerrum introduced the basic ideas for describing solutions where this condition is not satisfied. The situation will be clearer if we have in view some numerical values of the particular distance, which we can denote by b, for which $q^2/Kb = 2kT$ in different solvents at room temperature. Table 2 gives values for different types of ions in water and in ethyl alcohol ($K = 25$).

Table 2

	b in water	b in ethyl alcohol
Di-divalent	$13{\cdot}6 \times 10^{-8}$ cm.	$43{\cdot}2 \times 10^{-8}$ cm.
Uni-divalent	7·8	21·6
Uni-univalent	3·4	10·8

When in any solution a positive and a negative ion are at a distance apart less than the appropriate value of b, their Brownian motion will no longer be nearly independent; a certain time will usually elapse before they receive the necessary energy to move apart. Consider once more equations (7) and (8) with reference to a positive central ion. In Debye's treatment the values of equations (6) and (7) are nearly equal, both being near unity. For smaller distance from the central ion, on the other hand, equation (6) becomes small compared with equation (7). Suppose now that round each positive ion in the solution we draw a sphere of radius roughly equal to b. Within any sphere we may find either (1) no other ion, (2) one negative ion, (3) one or more positive ions, or (4) more than one negative ion; but in dilute solution cases (3) and (4) are extremely improbable, and we are left with the possibilities (1) and (2) only. At the point B in figure 1, where $r = a$, the field is due to the central ion only. For values of r slightly greater than a the potential will be approximately q/Kr, and the number of associated ions can be estimated by substituting in equation (7) $V_- = -q^2/Kr$. The probability that an ion of opposite sign lies inside a spherical shell of volume $4\pi r^2 dr$ is then

$$4\pi r^2 n \exp(2b/r)\, dr, \qquad \text{.....(29),}$$

where n is the concentration of the positive or negative ions. The probability that an ion of opposite sign lies within a sphere of radius b surrounding the central ion is obtained by integrating equation (29) from $r = a$ to $r = b$. Bjerrum's proposal[2] was to regard these ions, separated by distances less than b, as associated ion pairs, and to apply the Debye-Hückel theory to all the remaining ions. When a small potential

difference is applied to a solution, although the positive and negative members of each associated ion pair tend to move in opposite directions, they will be unable to separate until they receive the necessary energy from the solvent molecules. (In fields of ordinary low intensity there is no question of their receiving the necessary energy from the field itself; the field merely gives a bias to the Brownian motion.) Accordingly, each pair of associated ions fails to contribute to the electric current. Since equation (29) contains the concentration n, the proportion of associated ions will increase with the concentration, just as do the neutral molecules in the Arrhenius theory. The solutions will therefore deviate from the limiting Debye-Hückel laws in the observed direction. But in integrating equation (29) from $r=a$ to $r=b$, it was not clear why precisely the value b should be chosen as the upper limit for dividing the free ions from the associated ions, as was done by Bjerrum. Fuoss has shown how the choice of an arbitrary critical radius can be avoided.

In this treatment by Fuoss[3], *every* ion is paired with some one ion of opposite sign, and the short-range pairs will correspond to Bjerrum's associated ions. A positive and a negative ion, the centre of which lies at a distance between r and $r+dr$ from the positive ion, are defined to be an ion pair provided that no other unpaired negative ion lies within a sphere of radius r drawn round the positive ion. To visualize the meaning of this definition, suppose we surround each positive ion with a sphere of radius r, where r is at first equal to $a/2$, the average radius of the ions. We then let r for all ions increase uniformly, and as r increases we count the pairs formed. As soon as the circle round a given positive ion passes through the centre of a negative ion which has not been previously counted, those two ions are counted as a pair, associated with that particular value of r, and we no longer consider those two ions. In this way we eventually find a partner for each of the positive ions.

Let the centre of a particular positive ion be taken as origin, and let $G(r)\,dr$ be the probability that this ion forms a pair with one of the negative ions at a distance between r and $r+dr$, there being no unpaired negative ion nearer than a distance r. Then $G(r)$ will be proportional to n, to $4\pi r^2 dr/v$, to the Boltzmann factor

$$\exp(q^2/KrkT),$$

and finally to the probability $f(r)$ that an unpaired negative ion is not already present in the volume $\frac{4}{3}\pi r^3$ which is given by

$$f(r)=1-\int G(x)\,dx \qquad \text{.....(30).}$$

Combining these statements

$$G(r)\,dr=4\pi n r^2 dr\, e^{2b/r}\left\{1-\int_a^r G(x)\,dx\right\} \qquad \text{.....(31).}$$

For values of r little greater than a the values of $G(r)$ must agree with equation (29). With this boundary condition the solution of equation (31) gives

$$G(r)=4\pi n r^2 \exp\left\{\frac{2b}{r}-4\pi n\int_a^r x^2 e^{2b/x}\,dx\right\} \qquad \text{.....(32).}$$

A particular example of this function is plotted in figure 6; this was calculated by Fuoss for a concentration of 10^{-4} mol./l. of ions whose distance of closest approach is $4{\cdot}6 \times 10^{-8}$ cm.; as the curve depends on the value of q^2/K it applies equally to a di-divalent salt in water or to a uni-univalent salt in a solvent of dielectric constant about 20. At this concentration the thickness of the ionic atmosphere l is 150×10^{-8} cm., corresponding roughly to the position of the point E. The distribution is seen to have a deep minimum, separating the short-range pairs from the rest of the ions. At these low concentrations this minimum occurs very nearly at $r = b$; hence the short-range pairs correspond closely to the associated ions of Bjerrum. In figure 6 the short-range pairs are relatively few in number; but as the concentration γ is raised, the points A and C move upwards in proportion to γ, while the points D, E and F move relatively slowly[8] to the left and upwards in proportion to

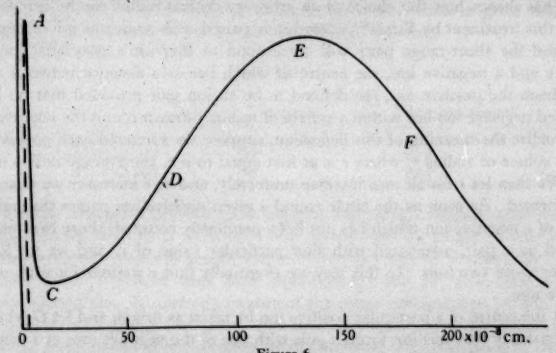


Figure 6

$\gamma^{\frac{1}{2}}$. Consequently at higher concentrations the fraction of associated ions increases and causes a progressive deviation from the limiting laws of dilute solution. The equivalent conductivity, for example, will fall more rapidly than the Onsager straight line, as in figure 5.

If α is the proportion of free ions, and $(1-\alpha)$ that of the short-range pairs, the equilibrium between them is determined by the mass action law[1], the Debye-Hückel theory being applied to the free ions. The average lifetime of an associated ion pair will be a very small fraction of a second; and the portion of the curve near C in figure 6 corresponds to ions momentarily in the transition stages between the free and the associated states.

In figure 6 the minimum occurs at about 14×10^{-8} cm. in agreement with the first value of b given in table 2. It will be seen from this table that aqueous solutions of uni-univalent substances appear to present a special case since a may be larger than b; species of ions for which a is less than b will behave differently from species for which a is greater than b. This prediction, however, is not fulfilled, for the reason

that in aqueous solution there are no ions for which the distance of closest approach is sufficiently small. In the solid crystal ions like Li^+ and F^- have small radii, but in water their position of minimum potential energy must be when there are one or more water molecules between the ions. In aqueous solution ions do not form short-range pairs. This failure is because the dielectric constant of water is so large; for a given value of a, however, there will always be a certain smaller value of K such that b becomes greater than a. Following up this idea, Fuoss and Kraus[9] took an aqueous solution and progressively diminished the dielectric constant of the solvent by mixing dioxane with the water (the dielectric constant of dioxane being only 2·2). Starting with a completely dissociated solution, and measuring the conductivities, they found that, as the value of the dielectric constant fell, association of the ions set in and increased at exactly the rate predicted by the theory, assuming only pure electrostatic forces to be operative.

In solvents of low dielectric constant, with increasing concentration of solute the equivalent conductivity, instead of showing a steady decrease, passes through a minimum value and rises again. In explanation of this phenomenon Fuoss and Kraus[10] have pointed out that the formation of short-range pairs will be succeeded by the formation of a small number of ion triplets, by the association of an ion with two ions of opposite sign, giving both $(+\,-\,+)$ and $(-\,+\,-)$. In the investigation of ion pairs, described above, these configurations were disregarded, but the number of ion triplets may be calculated by the same method as was used for short-range pairs. At high concentrations the presence of a small proportion of ion triplets becomes important, because a free ion incorporated into such a triplet does not lose its conductivity, as it does when it becomes a member of a neutral short-range pair. For the solutions which have been studied the trend of the conductivity is in satisfactory agreement with the calculated behaviour.

§ 8. CONDUCTIVITY IN FIELDS OF HIGH INTENSITY

We shall now leave the subject of associated ions and return to the conductivity of dilute aqueous solutions, where the presence of the ionic atmosphere is mainly responsible for the difference between Λ and Λ_∞. A new phenomenon was discovered in 1924 by Wien[11] in measuring the conductivity of solutions in very high fields. In these experiments a surge of electric current lasting for about a millionth of a second was sent through the solution. For low voltages the heating of the solution was in agreement with Ohm's law, but for field strengths of 100,000 v./cm., the heating was found to be much greater, indicating that the solution had a lower resistance and was passing a larger current. Later an oscillograph was used to follow the variation of the current and the potential during the pulse.

It is found that for sufficiently high fields the conductivity approaches a limiting value somewhat less than Λ_∞; which must mean that the ions are travelling with almost the mobility u_∞ with which they travel in pure solvent. It is evident that in these high fields both the retarding effects discussed in §6 must be greatly diminished

in value. It was mentioned there that the retarding field due to the asymmetry of the ionic atmosphere is proportional to the velocity with which the ion is moving; this is true for fields of ordinary low intensity, in which the ion moves with a velocity of the order of 0·001 cm./sec. But in Wien's experiments the ions travelled with a velocity of nearly a metre per second. In dilute solution, as soon as an ion acquires such a velocity it leaves its ionic atmosphere behind it. Consider, for example, a uni-univalent electrolyte at a concentration of 10^{-4} mol./l. The thickness of the ionic atmosphere $l = 3 \times 10^{-6}$ cm.; and the time of relaxation Θ is about 2×10^{-7} sec. During this time, in a field of 100,000 v./cm. an ion will have travelled more than 10^{-5} cm., or several times the thickness of its ionic atmosphere. In a less dilute solution the situation will be less extreme, but both the asymmetry and the electrophoretic effect will be smaller than in fields of low intensity. By successive approximations Joos and Blumentritt[12] obtained an expression for the increase in conductivity in the initial stages. For very high fields the asymmetry becomes negligible, but part of the electrophoretic effect remains. The higher the concentration, the greater is the field required to make the conductivity approach a limiting value.

For singly charged ions in dilute aqueous solution Λ differs so little from Λ_∞ that the total possible effect is small. Most of the experimental work has accordingly been done with divalent and trivalent ions, or in solvents of lower dielectric constant, where an increase of several per cent can easily be obtained. Both in aqueous solution and in acetone the increment in the conductivity is in approximate agreement with the theory of Joos and Blumentritt.

§ 9. THE DISSOCIATION FIELD EFFECT

Wien and his co-workers have carried out similar investigations upon solutions containing large numbers of associated ions, and have found that the effect of strong fields is even more marked. Figure 7, for example, gives some results[13] for $CoCl_2$ dissolved in acetone ($K = 20$), where in a field of 300 kv./cm. the conductivity has already increased by as much as 40 per cent, and shows no sign of approaching a limiting value. Effects of this magnitude can be explained only by an increase in the number of free ions by the action of the applied field—in fact, by just the mechanism which in the days before Arrhenius was proposed for the conductivity in ordinary weak fields. As was pointed out in §7, in any solution where the number of free ions and the number of associated ion pairs are comparable, this equilibrium is maintained by a rapid interchange. The average lifetime of an associated pair being a small fraction of a second, the number of ions forming short-range pairs in the course of their Brownian motion is balanced by the number of pairs breaking up. Although in a weak field this equilibrium is not affected, we must expect a large change when the fall of potential is sufficiently rapid—when the potential energy of an ion in the field changes by an amount comparable to kT in a distance of, say, 10^{-7} cm. In a field of 100,000 v./cm., for example, the potential energy of an electron will change by 0·01 ev. in a distance of 10^{-7} cm.; in weaker fields than this we may expect an easily measurable effect.

The action of such a field will be to cause the members of some ion pairs to move apart, and the members of other ion pairs to move closer together. If these numbers were equal, the degree of dissociation would not be changed. But, just as electrical dipoles become orientated in any external field, so will the members of ion pairs suffer an analogous constraint. In any solution, within a small fraction of a second after the application of a potential difference, there is no longer a uniform random distribution, but for both short- and long-range ion pairs there is a greater probability of finding the positive member on the side nearest the negative electrode, and the negative member on the side nearest the positive electrode[14]. With this spatial distribution the further action of the field upon the positive and negative members of ion pairs will clearly be to separate more than it brings together.

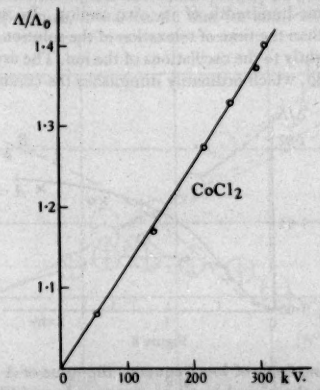


Figure 7. Abscissae give the field intensities in kilovolts per cm.

Onsager[15] has investigated the action of the field for the case where the number of free ions is small and the ionic atmosphere unimportant. He showed that the initial equilibrium between associated ions and free ions will be quickly replaced by another, in which the relative numbers of free ions is greater. The time required to change over from the initial to the new equilibrium should be less than 10^{-7} sec. The electrical conductivity of the solution, being proportional to the number of free ions present, will rise in the same ratio as the number of free ions is increased. The ratio R between the new dissociation constant and the old is given by

$$R = 1 + b + \frac{b^2}{3} + \frac{b^3}{18} + \ldots \qquad \ldots\ldots(33),$$

where for uni-univalent electrolytes $b = 9{\cdot}6\, X/KT^2$, X being the intensity of the field in volts per cm. The predicted magnitudes are quite large enough to account for results like those of figure 7.

§ 10. CONDUCTIVITY IN ALTERNATING FIELDS OF HIGH FREQUENCY

When Debye introduced the idea of the time of relaxation of the ionic atmosphere, it was clear that under alternating fields of sufficiently high frequency deviations from the normal conductivity should exist. When an alternating field is applied to a solution, each ion, apart from its random Brownian motion, will oscillate to and fro, the positive and negative ions moving always in opposite directions. If the field is not abnormally strong, the amplitude of this oscillation will become extremely small under high frequency. Thus in 10^{-7} sec., for example, with an ionic velocity of 10^{-3} cm./sec. the ion will not move more than 10^{-10} cm. If then a field of frequency 10^{7} is applied to a solution, each ion will oscillate through a distance less than one-hundredth of its own radius. In weak alternating fields whose period is less than the time of relaxation of the solution, the ionic atmosphere will respond only slightly to the oscillations of the ion. The asymmetry will be small and the retarding field, which ordinarily diminishes the conductivity, will be weak.

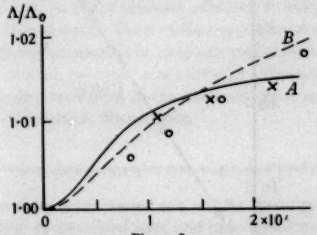


Figure 8

Although in alternating fields of low frequency the value of Λ is independent of the frequency, as soon as the period becomes comparable with the time of relaxation, the conductivity should increase and tend towards a limiting value. (This limiting value is not Λ_∞, since the electrophoretic effect is still present.) The magnitude of the change was calculated in 1928 by Debye and Falkenhagen[16], taking into account the fact that with increasing frequency the ionic motion gets more and more out of phase with the field. For the usual reasons a larger increase in conductivity is expected for highly charged ions than for singly charged; and this fact was used for confirming the existence of the predicted effects, comparisons being made between dissimilar solutions which had the same conductivity in steady (and in low-frequency) fields[17]. Figure 8 shows an increase of between 1 and 2 per cent in the conductivity of $MgSO_4$ in water, relative to KCl at a frequency of 2×10^7 c./sec.[18]. The crosses agree fairly well with the theoretical curve A; the circles and the theoretical curve B are for a concentration twice as great. Those solutions which show the abnormally large increment in intense fields would not be expected to show anything but the normal increment in alternating fields of high frequency; this was

verified by Whitmore[19]. The effect of temperature on the high-frequency increment has been investigated by Geest[20], and found to be in satisfactory agreement with the theory.

In all these investigations the measurements were relative, by comparison of one type of solution with another. Before any absolute measurements of conductivity at these frequencies could be made, it was necessary to have accurate data for the effect of these high-frequency fields upon the solvent itself. This work is referred to below in § 18.

§ 11. THE VISCOSITY OF IONIC SOLUTIONS

In measurements made during the nineteenth century on the viscosity of aqueous solutions at room temperature, it was found that for some salts the viscosity was

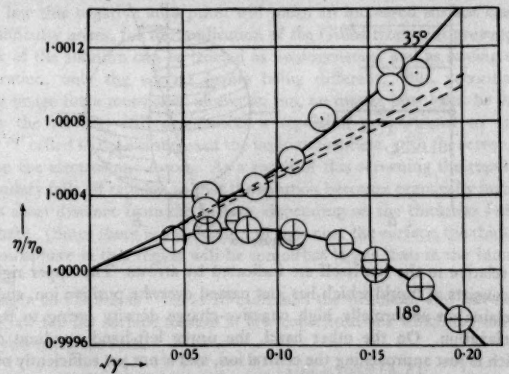


Figure 9. Relative viscosity of potassium chloride solutions at 18° and 35° C. (The dotted lines represent the theoretical limiting slopes.)

greater than that of pure water, and for others was less. In 1929 Jones and Dole[21], studying the viscosity of extremely dilute aqueous solutions, found that their results could be written in the form

$$\eta = \eta_0 (1 + A\sqrt{\gamma} + B\gamma) \qquad \ldots\ldots(34),$$

where η_0 is the viscosity of pure solvent. The $A\sqrt{\gamma}$ term becomes predominant at low concentrations, and they suggested that this term, like the $\sqrt{\gamma}$ terms which occur in the conductivity, etc., might be due to the Debye-Hückel forces. This suggestion has proved correct; the forces due to the ionic atmosphere, described below, cause an increment in the viscosity proportional to $\sqrt{\gamma}$. Many common salts, however, were known to give dilute aqueous solutions with viscosity smaller than that of water; this might be due to the fact that measurements had not yet been made at sufficiently low concentrations. Joy and Wolfenden[22] accordingly in-

vestigated some of these solutions, carrying their measurements to lower concentrations, and found that at great dilutions the value of the viscosity rose above the value of pure water, as shown by the lower curve in figure 9. These results can be expressed in the form of equation (34) by giving a negative value to B. At rather higher temperatures the value of B is no longer negative, as shown by the upper curve in figure 9. This B coefficient will be discussed in § 14; here we are interested in the interpretation of the $A\sqrt{\gamma}$ term.

Consider a liquid flowing parallel to the x axis, there being a velocity gradient parallel to the z axis, as indicated in figure 10*a*. We fix attention, as usual, on a particular positive ion, surrounded by its negative ionic atmosphere. The upper half of this atmosphere is moving faster than the ion, and the lower half is moving slower than the ion itself; hence the atmosphere is somewhat deformed. In figure 10*b* the

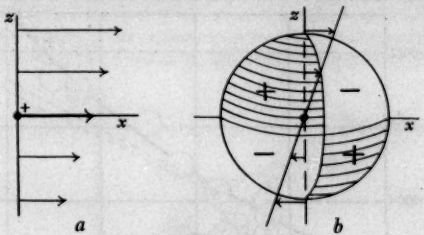


Figure 10

velocities relative to the ion itself are indicated by arrows. The upper right-hand quadrant consists of liquid which has just passed over the positive ion, and consequently retains an abnormally high negative-charge density owing to the finite time of relaxation. On the other hand, the upper left-hand quadrant contains liquid which is just approaching the central ion, and is not yet sufficiently negative; it is marked positive in the diagram. In the same way, the two lower quadrants are respectively positive and negative. It is clear that the electrical charges in the upper quadrants will tend to carry the central ion forward in the x direction; the force will be in the same direction as the viscous drag of the solvent. Between the central ion and the charge in the lower quadrants there is a similar drag. Falkenhagen[23] determined the form of the charge distribution, and calculating the total shearing force transmitted through unit area of the liquid at right angles to the velocity gradient, he found that the viscosity should take the form

$$\eta = \eta_0 (1 + A\sqrt{\gamma}) \qquad \text{.....(35).}$$

For binary electrolytes with ions of equal mobility the value of A is given by

$$A = \frac{0{\cdot}512\, z^2}{u\eta\, (KT)^{\frac{1}{2}}} \qquad \text{.....(36).}$$

As is seen in figure 9, the observed values agree with the square-root law only as far as the limiting slope is concerned.

§ 12. SURFACE TENSION

Another property of electrolytic solutions where the ionic atmosphere must be taken into account is their surface tension. It is well known that when an electric charge is situated in a vacuum near the surface of a body, whether conductor or insulator, the charge is attracted towards the surface of this body by an electrical image force; for a plane surface the intensity of this force varies inversely as the square of the distance from the surface. Electrical image forces are also experienced by charges situated *within* a dielectric medium; for when a dielectric, such as a solvent, is in contact with a medium of lower dielectric constant, such as air, any charge situated within the solvent will be repelled from this boundary, the force again being inversely proportional to the distance from the boundary. In any solution, therefore, all the ions will be repelled from the surface; and according to Gibbs's law this negative adsorption will mean an increased surface energy. But here a difficulty arises, for the application of the Gibbs treatment presupposes that the bulk of the solution can be treated as homogeneous, and as having a uniform concentration, only the surface layers being different. But, according to the ordinary image force mentioned above, an ion, no matter how great be its distance x below the surface, still experiences a repulsion proportional to $1/x^2$. Here Wagner[24] called in the assistance of the ionic atmosphere, with the screening which exerts on the electrostatic forces. As a result of this screening the repulsion from the boundary falls off rapidly, so that the solution becomes practically homogeneous within a short distance from the surface, depending on the thickness l of the ionic atmosphere. (Since there is a deficiency of ions near the surface, the thickness of an ionic atmosphere in this region will be somewhat larger than in the interior of the solution.)

Onsager and Samaras[25] have re-examined the problem, and have deduced a limiting law for the surface tension at low concentrations which for uni-univalent electrolytes takes the form

$$\sigma = (\sigma_0 + 79{\cdot}52/K)\,\gamma \log_{10} \frac{1{\cdot}14 \times 10^{-13}\,(KT)^3}{\gamma} \qquad \text{.....(37).}$$

The predicted increment in the surface tension for water at room temperature amounts to one part in a thousand at $\frac{1}{10}$-M. solution. This agrees fairly well with the observed values; but Jones and Ray[26] have recently reported that at extreme dilutions (below 0·006 M.) the surface tension of several ionic solutions is slightly *less* than that of water.

§ 13. THE SOLVATION OF IONS

In accounting for various properties of ionic solutions in the foregoing sections, it will be seen that each interpretation has been scarcely more than a bald attempt to predict the behaviour of an assembly of charged particles. The results are applicable, and indeed have been applied, to ions in the interior of stars. In applying the same results to electrolytic solutions, the properties of the particular species of ions have been almost completely disregarded, and no account has been taken of the molecular

nature of the solvent. We have seen how the ideas of ionic atmosphere and ionic association were found to provide a connecting link between diverse properties of the same solution. In no other aspect of electrolytes has such steady progress been maintained over a number of years. Whereas in gases and vapours the nature and structure of atomic ions are now known in considerable detail for nearly every element in the periodic table, relatively little has been done to describe the status of the same ions when situated in a solvent, or the processes by which they are removed from solution in electrolysis. Prior to 1920 it was thought necessary to provide a mechanism whereby neutral substances should become dissociated into ions on solution, and chemical combination between ion and solvent was proposed; but as soon as the view was accepted that most crystalline salts already consist of positive and negative ions, this became unnecessary. To detach an ion from such a crystal a certain amount of work must be done against the electrostatic forces which hold it in place; in a vacuum, or in air, these forces are large, but in a solvent of dielectric constant K they are some K times smaller. The chief progress has come from the recognition since 1912 that the large dielectric constants of the most common solvents arise from the fact that these liquids consist of polar molecules, and that in any electrostatic field the tendency for these dipoles to become aligned by the field is opposed by their thermal motion. More recently the view has been put forward[27] that these liquids possess in addition certain traces of local structure, i.e. when a substance such as ice melts, the structure does not entirely disappear, but the relation between a molecule and its neighbours is still to some extent determined by the interaction between the electrical dipoles[28]. Formerly, many of the properties of water were attributed to the fact that at room temperature, and below, a certain fraction of the molecules were polymerized to form $(H_2O)_2$ and $(H_2O)_3$, which polymers tended to break down as the temperature was raised. This statement is now commonly replaced by the assertion that the local structure in the liquid becomes progressively more disordered at higher temperatures. Recently the Raman spectra of ice, liquid H_2O, and deuterium oxide D_2O, have been investigated, to obtain information as to the relation between adjacent molecules. The alteration in the intensities of the Raman lines with temperature is found to be consistent with the concept of small structural changes in the liquid[29].

To describe the interaction between a particular species of ion and a particular solvent—what is known as the solvation of the ion—we have first to describe the effect of the electric charge of the ion upon the polar solvent molecules. Towards a positive ion a small number of solvent dipoles will present their negative ends, while towards a negative ion they will turn their positive ends. To describe the discharge of an ion from solution in electrolysis, we have to describe the removal of the ion from the solvent dipoles. In §2 we recognized what a large number of the familiar atomic ions have the inert-rare-gas structure. Add to this the fact that the water molecule itself has the same number of electrons as the neon atom, and it appears that in aqueous solution the electrostatic interaction between ionic charge and solvent dipoles is not likely to be accompanied by exchange forces. Some results for the effect of dissolved ions upon the Raman spectrum of water have already been

obtained[30]. In the theoretical approach a difficulty arises from the fact that in the neighbourhood of an ion the intensity of its field approaches a million v./cm., and we cannot predict the properties of a liquid subject to such an intense field. In a certain volume round each ion the solvent will have neither its usual dielectric properties nor its usual mechanical properties. In §11 attention was drawn to the fact that for some aqueous solutions the viscosity at room temperature is less than that of pure water, the coefficient B in equation (34) being negative. As early as 1900 this was attributed to a local depolymerization of the water by the ions[31]; this is to say, in a small volume round each ion the structure of the water receives the same alteration which it would have at a higher temperature. In agreement with this view we see from the upper curve of figure 9 that if the same ions of KCl are introduced into water at 35° C., they no longer reduce the viscosity, because now the local structure has already become everywhere more disordered by the thermal agitation.

§ 14. THE MOBILITY OF IONS

In the discussion of electrical conductivity in §6 it was pointed out that the Debye-Hückel theory deals with the difference between u and u_∞, but makes no attempt to say anything about the absolute value of u_∞ for any ion, nor how the value of u_∞ will vary with temperature, etc. We may now briefly consider the characteristic mobility of an ion in relation to the viscosity. We cannot predict whether Stokes's law should be strictly obeyed for particles so small as ions; but if it is obeyed, the value of u_∞ for each ion will by equation (23) be inversely proportional to the viscosity of the solvent. At 50° C. the viscosity of water has almost exactly *half* the value that it has at 18°. For some salts in aqueous solution the value of Λ_∞ at 50° is nearly *double* the value of 18° (so it appears that by these ions Stokes's law is obeyed); but for other salts the value is considerably less than double; for KCl, for example, the ratio is nearer 1·8 than 2·0; we may tentatively call this behaviour anomalous. Now Cox and Wolfenden[32] have drawn attention to the interesting fact that the particular salt solutions which show the greatest anomaly in this respect are those which have negative B-coefficients in the formula (34) for their viscosity at room temperature; which suggests a satisfactory explanation of this behaviour. The value of u_∞ depends, not on the properties of the main body of the liquid, but on the local properties of the liquid surrounding each moving ion at the moment. And we have seen that a large negative B-coefficient probably means that at room temperature the liquid round each ion has locally a structure similar to that which it would have at a higher temperature; we should therefore expect the value of Λ_∞ to vary with temperature much less rapidly than the value of the macroscopic viscosity η for the liquid as a whole. On the other hand, for ions which have positive B-coefficients at room temperature the behaviour should be more in accordance with Stokes's law. This is just the difference between the two types which is observed.

§ 15. PROTON AND DEUTERON TRANSITIONS

In another important but restricted field progress has been made in the interpretation of those ionic mobilities where a proton or a deuteron is concerned. Indeed, the whole status of the proton in solution, particularly in aqueous solution, has recently become one of the major problems of electrolytics, largely through the dominant role which it has been found to play in biophysics. In the water molecule the ten electrons circulate round the oxygen nucleus, and the two protons are embedded in the general electron cloud. The lines joining the protons to the oxygen nucleus are believed to make an angle of about 106° with one another[28], as in figure 11. If we remove one of these protons, the result is a negative hydroxyl ion, as in figure 11. In the NH_3 molecule there are three protons embedded in the electron cloud; similarly when a proton is introduced into water, there is no doubt that it takes up a position inside a water molecule to form an ion $(H_3O)^+$, analogous to the NH_3 molecule.

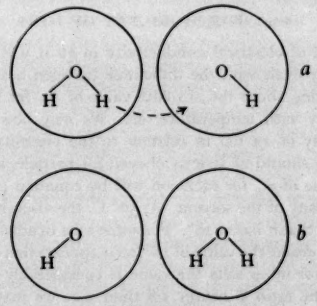


Figure 11

When the mobilities u_∞ for the common ions in aqueous solution were first determined, it was found that the values for the hydrogen and hydroxyl ions were exceptionally high. For all the common ions the values lie within a remarkably narrow range, but for the hydroxyl ion the value is nearly three times as great, and for the hydrogen ion nearly five times as great. It had long ago been suggested that this phenomenon might be due to some form of Grotthus chain conduction. The more detailed picture which we have of the H_2O molecule now enables us to visualize the process[28]. Let figure 11 *a* represent a hydroxyl ion and a water molecule side by side in the interior of liquid water. Similarly figure 11 *b* represents a water molecule and a hydroxyl ion side by side, a configuration which has of course exactly the same energy as that shown in figure 11 *a*. But obviously *a* can be converted into *b* by transferring a proton from the H_2O molecule to the $(OH)^-$ ion, as indicated by the arrow. When the proton has been transferred from left to right, it is as if the hydroxyl ion had moved bodily from right to left. If such transitions can take place

spontaneously, any applied field will give them a directional bias. Like any other ion, the hydroxyl ion can, and will, migrate bodily through the liquid, but the abnormally high mobility must be due to the protonic transitions. Similar considerations apply to the mobility of the $(H_3O)^+$ ion in water.

Support has been given to this interpretation by the recent discovery that a similar mechanism appears to be present when pure H_2SO_4 is used as solvent. Owing to the high viscosity of this liquid most ions have very low mobilities in it; but in a solution of barium or strontium bisulphate, it is found that the negative ion has a much higher mobility[33]. As in figure 11, consider a negative bisulphate ion $(HSO_4)^-$ and an adjacent H_2SO_4 molecule; it is clear that if a proton is transferred from the latter to the former, the molecule becomes a $(HSO_4)^-$ ion, and the ion becomes a solvent molecule. Consequently conduction by means of proton transitions is possible, just as for the hydroxyl ion in water. But in sulphuric acid the observed effect is much more pronounced, since the normal migration of ions is heavily suppressed by the high value of the viscosity.

What has been said of the proton in water applies also to the deuteron in heavy water. The viscosity of liquid D_2O at 25° C. being 23·2 per cent greater than for H_2O[34, 35], the mobility of all the common ions in D_2O is correspondingly lower, but the ion $(D_3O)^+$ has again a remarkably high mobility, which must be attributed to deuteron transitions. If in figure 11 every H is replaced by a D, the configurations *a* and *b* again provide two states of exactly the same energy. But if we go on next to consider a mixture of H_2O and D_2O, we see that a variety of configurations will occur, some of which will be unsymmetrical; for example, for the deuteroxyl ion

$$(OD)^- + HDO \rightleftharpoons D_2O + (OH)^-,$$

and for the positive ion

$$(H_2DO)^+ + HDO \rightleftharpoons H_2O + (HD_2O)^+.$$

In such cases we do not have alternative configurations of the same energy, and consequently the occurrence of the transition has a lower probability. This explanation has been proposed for the experimental results shown in figure 12 for the conductivity of HCl and DCl in the mixed solvent from 100 per cent H_2O to 100 per cent D_2O[35]. The variation of the conductivity is far from linear; in the mixed solvent the curve lies everywhere below the straight line. By following up such comparisons between the proton and deuteron it should be possible to throw light upon the behaviour of hydrogen ions in various solvents. Already the discovery that H_2 and D_2 can be fractionated by means of electrolysis has stimulated a fresh inquiry into the processes taking place in the discharge of ions at the surface of a metal electrode[36].

§ 16. IONIC DISSOCIATION OF MOLECULES

It was with aqueous solutions of substances like acetic acid that the equations of Ostwald and Arrhenius had their conspicuous success. Now that the idea of short-range ion pairs has become current, some doubt may be felt as to the relation between these associated ions and true neutral molecules. In a neutral CH_3COOH

molecule the loosely bound proton is embedded within one of the oxygen atoms, like the protons in H_2O. Such a structure can only be regarded as a true molecule. Consider now any *R*OOH acid in solution. When ionic dissociation of the molecule takes place, the first step is the transference of the loosely bound proton to an adjacent solvent molecule; thus in aqueous solution, for example,

$$R\text{OOH} + H_2O \rightarrow (R\text{OO})^- + (H_3O)^+ \qquad \text{.....(38).}$$

For this initial step a certain amount of work must be done, depending upon both the particular solute and solvent molecules involved. Separation of these adjacent positive and negative ions will complete the dissociation. For this second step a further amount of work must be done against the electrostatic forces of attraction. The sum of these two contributions constitutes the dissociation energy of the molecule in solution. If at any time the ions recombine to form a neutral molecule, this potential energy is reliberated. In spite of this, recombination of ions is often accompanied by an absorption instead of an evolution of heat[37]. This is due to the

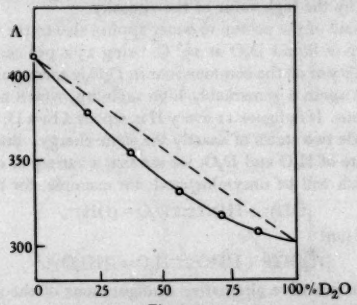


Figure 12

fact that, when the electric charges recombine, the surrounding polar medium is subjected to a decreasing electric field, which gives rise to a cooling effect[38].

Recently systematic measurements of the ionic dissociation of a dozen different substances in aqueous solution have been made over a wide range of temperature from 0° to 60°[37]. The dissociation of acetic acid has also been studied in water mixed with dioxane[39] to reduce the dielectric constant, as in §7. All the substances investigated show a similar behaviour. Whereas in any gas molecular dissociation always increases steadily with rising temperature, for each of these substances in solution the degree of dissociation reaches a maximum at a characteristic temperature, θ, and falls off again. In accordance with the well-known principle, the heat of dissociation becomes zero at θ and changes sign. For temperatures below θ recombination of ions is accompanied by an evolution of heat, and for temperatures above θ by an absorption of heat. This behaviour is due to the fact that the electrostatic forces vary rapidly with temperature, while the non-electrostatic forces in the reaction (38) are nearly independent of temperature[38].

We may pass on now to the electrical conductivity of aqueous solutions of *R*OOH acids. When measurements were made in very intense fields, as in §8, large increases in conductivity were found[40]. For acetic acid an increase of 16 per cent was found in a field of less than 300 kv./cm. Since the field will have no direct action on the dissociation of neutral molecules, it might be thought that the observed effect indicated that a large part of the solute is in the form of short-range ion pairs; this conclusion, however, appears to be unnecessary and incorrect. (There could be large numbers of associated ion pairs, if the radius of the solvated hydrogen ion were sufficiently small; the properties of dilute solutions of HCl are, however, similar to those of KCl, showing no trace of association.) In an intense field the usual equilibrium between free ions and acetic acid molecules is clearly replaced by another equilibrium in which the proportion of free ions is greater[15]. The lifetime of a CH_3COOH molecule in aqueous solution being a small fraction of a second, at any moment some ions will be in the act of moving apart. An intense field will rapidly separate these, and the neutral molecules will act as a reservoir, supplying further ion pairs, which in turn are separated by the field, until the new equilibrium is rapidly set up between free ions and neutral molecules.

§ 17. THE VOLUME AND COMPRESSIBILITY OF IONIC SOLUTIONS

In discussing the viscosity of a dilute solution, we saw that the effect consists of two parts. One part arises from the interaction of each positive and negative ion with the solvent—the solvation of the ions; this effect, depending upon the number of each species per unit volume, gives rise to the $B\gamma$ term. Superimposed on this is an effect arising from the interionic forces, and giving the $A\sqrt{\gamma}$ term. Several other properties of ionic solutions consist of two analogous parts, one arising from the solvation of the individual ions, and the other, proportional to $\sqrt{\gamma}$, arising from the ionic atmosphere. In the concluding sections of this chapter we shall briefly consider some properties connected with the polar molecules of the solvent.

For solutions containing only neutral molecules the volume of the solution is in some cases greater and in others less than the sum of the initial volumes of solvent and solute. For ionic solutes the diminution in volume is very marked owing to the electrostatic attraction between the ionic charge and the oriented dipoles. For various species of positive and negative ions in dilute aqueous solution the effect is an additative one[28]. Superimposed on this is an effect arising from the ionic atmosphere. In §3 we saw that if, by moving a semi-permeable membrane, we confine the solute in a smaller volume, the ions will fall to a state of lower electrostatic potential energy. Now there is obviously a simpler way in which the volume occupied by the solute may be slightly diminished, namely, by diminishing the volume of the whole solution, as we do in measuring the compressibility of a solution. As before, the solute will fall to a state of lower electrostatic energy, by an amount depending on the concentration. If this were the only effect, it would be obvious in which direction the ionic atmosphere would modify the volume and compressibility of the solution. But there is another effect, arising from the change

in the dielectric constant of the medium. When a polar liquid is compressed, the value of its dielectric constant increases. Now an increase in K means weaker electrostatic forces between the ions, and therefore a movement towards a state of higher potential energy. In this respect the presence of the ionic atmosphere acts in the opposite direction to the former. From the Debye-Hückel free energy Redlich and Rosenfeld[41] derived the expression giving the change of apparent molal volume with concentration, and showed that both the above effects, being determined by the thickness l of the ionic atmosphere, will vary as the square-root of the concentration; the question is which is the larger of the two. For water at room temperature the increase of K with pressure is known, and the magnitude is such as to make the second effect about three times as large as the former. Consequently with increasing concentration in all dilute ionic solutions the apparent molal volume should increase, the change being more rapid for doubly and trebly charged ions. The observed behaviour is in agreement with this, but only qualitatively[42]; a similar amount of agreement is found for the apparent molal compressibilities[42].

§ 18. THE DIELECTRIC CONSTANTS OF IONIC SOLUTIONS

As was mentioned in § 13 the high dielectric constant of a polar liquid is due to the partial orientation of the dipoles in any applied field. Now in the neighbourhood of any ion the radial field is far more intense than any external field applied to a liquid; consequently the dipoles adjacent to any ion will be less free to turn than they are in pure solvent. Until 1928 it was thought that owing to this saturation effect the dielectric constant of any ionic solution would be less than that of pure solvent[43], this diminution being present in fields of all frequencies. In an ionic solution, however, not only the solvent molecules but also the ions respond to the field, and it was recognized by Debye and Falkenhagen[16] that this leads to an increase in the dielectric constant, proportional to $\sqrt{\gamma}$, and depending on the charges borne by the ions. It was calculated that in a $\frac{1}{100}$-M. aqueous solution the increment should amount to 3·03 per cent for $MgSO_4$, and to 0·38 per cent for KCl; the experimental results in low-frequency fields are found to agree with these values within the experimental error[44].

Recently measurements have also been made in alternating fields of high frequency, where the period is comparable with the time of relaxation of the ionic atmosphere. There being scarcely any data for the pure solvents, each investigation deals with the solvent itself, before going on to solutions. For any polar liquid, as Debye pointed out[43], under high-frequency fields there arises a difference in phase between the electric polarization and the field, owing to the fact that the polar molecules themselves possess a time of relaxation. For each liquid this characteristic time is less than 10^{-10} sec., much shorter than the periods of 10^{-7} sec. which are used in the investigation of solutions. Nevertheless, even for frequencies as low as 10^7 c./sec. the difference in phase between the molecular dipoles and the field causes a large absorption of energy from the field. The resulting heating of the solvent must be allowed for in all absolute measurements of the conductivity and

dielectric constant of solutions at these frequencies, since the heating may cause an increment in conductivity several times larger than the effect looked for.

In any ionic solution under fields approaching 10^7 c./sec. the value of K begins to depend upon the frequency for the same reason that the electrical conductivity depends on the frequency. It was shown by Debye and Falkenhagen that, as soon as the period becomes comparable with the time of relaxation Θ of the ionic atmosphere, the increment in K begins to disappear. The experimental results for various species of ions are in good agreement with the theory; the value of K returns towards the value characteristic of the solvent itself at a rate depending upon the concentration.

Leaving now the interionic forces, there remains only the solvation of the ions, with the so-called saturation effect mentioned above. This effect, being proportional to γ, while the interionic effect is proportional to $\sqrt{\gamma}$, should introduce a diminution in K at higher concentrations. But even in solutions up to 2-M. recent measurements failed to disclose any definite deviation; the saturation effect was apparently at least ten times smaller than had previously been supposed[45].

REFERENCES

(1) Debye and Hückel. *Phys. Z.* **24**, 185 (1923).
(2) Bjerrum. *K. danske vidensk. Selsk. Skr.* **7** (1926).
(3) Fuoss. *Trans. Faraday Soc.* **30**, 967 (1934).
(4) Schreiner and Frivold. *Z. phys. Chem.* A, **124**, 1 (1926).
(5) Onsager. *Phys. Z.* **28**, 277 (1927).
(6) Sementschenko and Serpinsky. *Z. phys. Chem.* A, **167**, 197 (1933).
(7) Frazer and Hartley. *Proc. roy. Soc.* A, **109**, 351 (1925).
(8) Fuoss. *Chem. Rev.* **17**, 32 (1935).
(9) Fuoss and Kraus. *J. Amer. chem. Soc.* **55**, 1019 (1933).
(10) Fuoss and Kraus. *J. Amer. chem. Soc.* **55**, 2387 (1933).
(11) Wien. *Ann. Phys., Lpz.*, **83**, 327 (1927).
(12) Joos and Blumentritt. *Phys. Z.* **28**, 836 (1927).
Blumentritt. *Ann. Phys., Lpz.*, **85**, 812 (1928).
(13) Bauer. *Ann. Phys., Lpz.*, **6**, 253 (1930).
(14) Fuoss. *Phys. Z.* **35**, 60 (1934).
(15) Onsager. *J. chem. Phys.* **2**, 599 (1934).
(16) Debye and Falkenhagen. *Phys. Z.* **29**, 401 (1928).
(17) Sack. *Phys. Z.* **29**, 627 (1928).
Arnold and Williams. *J. Amer. chem. Soc.* **58**, 2620 (1936).
(18) Deubner. *Phys. Z.* **33**, 223 (1932).
(19) Whitmore. *Phys. Z.* **34**, 649 (1933).
(20) Geest. *Phys. Z.* **34**, 660 (1933).
(21) Jones and Dole. *J. Amer. chem. Soc.* **51**, 2950 (1929).
(22) Joy and Wolfenden. *Proc. roy. Soc.* **134**, 424 (1931).
(23) Falkenhagen and Dole. *Phys. Z.* **30**, 611 (1929).
Falkenhagen and Vernon. *Phys. Z.* **33**, 140 (1932).
(24) Wagner. *Phys. Z.* **25**, 474 (1924).
(25) Onsager and Samaras. *J. chem. Phys.* **2**, 528 (1934).
(26) Jones and Ray. *J. Amer. chem. Soc.* **59**, 195 (1937).
(27) Stewart. *Phys. Rev.* **37**, 9 (1931).
(28) Bernal and Fowler. *J. chem. Phys.* **1**, 515 (1933).
(29) Cross, Burnham and Leighton. *J. Amer. chem. Soc.* **59**, 1134 (1937).

(30) Hibben. *J. chem. Phys.* **5**, 166 (1937).
(31) Sutherland. *Phil. Mag.* **50**, 481 (1900).
(32) Cox and Wolfenden. *Proc. roy. Soc.* **145**, 487 (1934).
(33) Hammett and Lowenheim. *J. Amer. chem. Soc.* **56**, 2620 (1934).
Hantzch. *S.B. Akad. Wiss. Wien*, **63**, 1794 (1930).
(34) Lewis and MacDonald. *J. Amer. chem. Soc.* **55**, 4730 (1933).
(35) Baker and La Mer. *J. chem. Phys.* **3**, 408 (1935).
(36) Urey and Teal. *Rev. Mod. Phys.* **7**, 44 (1935).
Farkas. *Trans. Faraday Soc.* **33**, 552 (1937).
(37) Harned and Embree. *J. Amer. chem. Soc.* **56**, 1050 (1934).
(38) Gurney. *Ions in Solution*, Chapter 13 (1936).
(39) Harned and Kazanjian. *J. Amer. chem. Soc.* **58**, 1912 (1936).
(40) Schiele. *Ann. Phys., Lpz.*, **13**, 811 (1932).
(41) Redlich and Rosenfeld. *Z. phys. Chem.* A, **155**, 65 (1931).
(42) Gucker. *Chem. Rev.* **13**, 116 (1933).
(43) Debye. *Polar Molecules.*
(44) Wien. *Ann. Phys., Lpz.*, **3**, 629 (1935).
Williams and Arnold. *Acta Physicochimica*, **3**, 629 (1935).
Mazee. *Phys. Z.* **37**, 914 (1936).
(45) Hackel. *Phys. Z.* **36**, 220 (1935).

PHYSICAL-OPTICAL INSTRUMENTS AND MATERIALS

By F. SIMEON, D.Sc., F.Inst.P.

§ 1. NEW MATERIAL

For many years a large range of glasses has been available so that, although the computer may have wished for an even more gradual change in properties, he has nevertheless by judicious selection been enabled to achieve many remarkable lens designs. While this has served the purpose of most instruments embodying telescopes, photographic lenses and the like, the case has been very different when regions of spectrum were in question outside the limits to which glass is transparent. Indeed, until just a few years ago the substances available for use within these further limits comprised only crystalline quartz, rock salt, calcite and fluorite, and of these calcite has a limit of transparency at about 2300 A. in the ultra-violet and is practically useless in the infra-red, while fluorite is only available in small pieces. The latter difficulty has now been overcome by the optician by joining pieces of suitable shape to make lenses and prisms of the required size. Such lenses would, in general, be used as components in achromatic combinations and are best used between components of some robust material such as quartz.

These naturally occurring substances have now been supplemented by materials such as fused silica and crystals of KCl, KBr, etc. Fused silica has been known for many years for its remarkable properties—very low expansion, ultra-violet transparency, high melting-point, chemical resistance, etc.—and has found many applications in mercury-vapour lamps, laboratory utensils, etc. From an optician's point of view, however, it has for long proved disappointing on account of the presence of innumerable small bubbles and heterogeneities which made themselves very evident in interferometers or in polarized light. It is now available in qualities free from these defects, and while it has not become popular for the prisms of spectrographs on account of its small dispersion, it has been successfully employed for interferometers of the Fabry-Perot type. The first such use was by Williams[1], who employed it in a spectro-heliograph for which it was peculiarly adapted on account of the properties mentioned above, the etalon being exposed to intense radiation from the sun. During these experiments the tenacity of the optical contact possible with two silica surfaces was noticed, and this led later to the realization of a reflexion echelon, the form of echelon originally conceived by Michelson. Not only were the etalon plates made of silica but the separating pieces were also, and when coupled by optical contact the whole etalon proved extremely stable both in use and in transit. The separators in this instance were only about 3 mm. thick, but more recently an etalon has been used and described by Jackson[2] in which the separation was no less than 10 cm. The permanence of this etalon can be realized

from the fact that it was assembled before silvering, and that this process was carried out by the evaporation method without detriment to the contact between the etalon plates and separators. The quality both of workmanship and material must evidently be of the highest in an instrument of this power, more particularly in the case of the plate nearer to the observer. Conditions with regard to the material itself are not so stringent in the case of the reflexion echelon mentioned[3], as the light does not pass through it, although the internal condition may not be without effect when a number of plates are assembled together. Suffice it to say that echelons with as many as 40 plates have been successfully constructed.

Considering now the infra-red end of the spectrum, the materials generally available for the production of prisms are quartz, fluorite and rock salt. With these, however, the practical long-wave limit is at about 17 μ., while the fundamental frequencies of very many substances correspond to still longer wave-lengths. Sylvine, a natural crystalline form of potassium chloride, has long been known to be transparent to about 21 μ.; but its occurrence is so rare and in such small crystals that the naturally occurring material cannot be seriously considered. Largely due to the work of Pohl in Germany, however, large single crystals of this substance, and of the kindred potassium bromide, transparent to 30 μ., have been produced from which serviceable prisms can be cut. The method of production appears to involve the fusion of very pure chemicals in a suitable furnace and the subsequent cooling under pressure at a controlled rate. While such crystals may have been produced primarily for the investigation of other crystalline properties, the spectroscopist is indebted to the producers for a means of greatly extending the range of his investigations. Not only so, but in the case of sylvine at least a further substance has been added to the short list of those available to computers of achromatic lenses for the ultra-violet and utilized to such purpose that a lens of focal length 1000 mm. employing components of quartz, fluorite and sylvine has been designed in which the chromatic aberration between 8000 and 1850 A. is only 4 mm.[4].

Another substance which has been produced in the same way but which is of interest in the Schumann region is lithium fluoride. This has also been produced by Pohl and more recently by Stockbarger[5]. It has been claimed that certain specimens of this material are transparent to shorter waves than the limit of transmission of Schumann's fluorite parts. The latter seem to have been of exceptional material, as his limit of wave-length does not appear to have been surpassed, even if it has been equalled, in more modern instruments, and this in spite of superior vacuum technique.

The substances which have just been considered have special uses and, though of great value scientifically, will appeal to a very limited class of users only. There are, however, certain other substances which are likely to have a much wider field of application. One such class is the synthetic resins, and of these one in particular—methyl methacrylate, to which has been given the trade name "Perspex"—is of special interest. It is perfectly clear and colourless and capable of taking a high degree of polish. It is lighter in weight and much less brittle than glass, so that it has obvious attractions for use in such instruments as field glasses. As, further, finished

parts can be produced by moulding, parts which are required in quantity should be low in cost. Much more work seems to be necessary to establish the mechanical and optical properties of this material, but should present hopes be justified there would appear to be an extensive field for its use.

Another interesting substance but with very different optical properties has also recently become available, viz. polaroid. In the process of manufacture, minute crystals are marshalled in regular sequence in a suspension which on drying has the property of transmitting only light which is plane-polarized. It is evident, therefore, that two pieces of such material can be so disposed as to extinguish light when they are placed in the beam, and although the extinction thus produced is not equal to that obtainable with polarizing prisms, there are applications in which a large beam of polarized light is required for which it is very valuable. When polarizing prisms only are available, lens systems must be used to concentrate the light on to the prisms and then extend the beam so as to illuminate the object under examination with subsequent reduction to pass through the second prism. This involves three or more lenses which must be free from double refraction and not strained in the mounting. Polarization by reflexion has also been largely employed, particularly in the class of instruments known as "strain-viewers", for the examination of industrial glassware. An interesting example of this use for large specimens is to be found in the Louvre[6] strainviewer. In this instrument the louvres are so disposed that the light is reflected from each at the polarizing angle and thence converges to the observing eyepiece which is provided with a wave-plate.

A further class of substances which should be mentioned is that of liquids. Hollow prisms are, of course, no novelty, and many instances can be found in the literature of the use of prisms of carbon disulphide, although one objection to the use of this substance is its volatility. The need on occasion for combining high dispersion and large aperture and the difficulty of obtaining sufficiently large pieces of material to make prisms, has led to the construction of large hollow prisms which, when suitably thermostated, can be used in spectrographs of good definition. One liquid which, on account of freedom from colour, high dispersion, fairly low volatility, etc., has become of value in this connexion is ethyl cinnamate.

Another liquid, which, although far from new in itself, has been the subject of considerable investigation, more particularly in France, is water. Its transparency to the ultra-violet has led to its being put forward as the prism material in spectrographs, and it has also been proposed as the component of an achromatic lens for the same region. The quartz-fluorite-water lens designed by Bricout[7], although not quite equal to the quartz-fluorite-sylvine lens computed by Perry is, nevertheless, a very great improvement upon a quartz-fluorite doublet.

But the many disadvantages attending the use of liquids for optical purposes make it likely that they will be avoided except in cases where similar performance cannot be obtained with the use of solid components only.

§ 2. INSTRUMENTS

Among the physical properties of substances there are some which permit of simple determination and which serve to measure the concentration of the substance, amount of impurity, etc. To mention two only—refractive index and rotatory power—we have properties to measure which instruments have been worked out by physicists and which have then been adopted by chemists and constantly used by them for certain classes of work, principally of a routine character. In this way they have passed into industry, and certain essential conditions have thus been found which an industrial instrument should satisfy. Among these may be mentioned robustness, reliability, simplicity of operation and rapidity of indication. Of recent years, a similar course has been followed in the case of spectrum analysis, although in this instance the instrument has tended to pass more directly from the academic to the industrial laboratory.

It is, of course, more than a century since the first spectroscope was constructed by Fraunhofer, and three-quarters since Kirchhoff and Bunsen observed the flame spectra of metals. Until less than twenty years ago, however, there were very few names connected with spectrum analysis, and while the work of these pioneers was very valuable it is not surprising that instrument design had kept fairly closely to classical lines. One point had indeed been firmly established, namely, the importance of the ultra-violet spectrum, and hence the necessity for a spectrograph dealing with this region, usually one with prisms and lenses of quartz. As might be expected, the earlier quartz spectrographs followed the visual spectrometer very closely, a camera replacing the eyepiece of the spectrometer telescope. Also, lenses were then made achromatic—either quartz-fluorite or quartz-calcite combinations—but with these, as might have been expected, there was considerable curvature of the focal plane, so that only a very limited region could be obtained in focus in the centre of the plate, and not only were adjustments of focus needed when the range was varied, but to alter this range the whole camera was moved about the axis of the prism table. A notable improvement was made in 1908 by the introduction by Twyman of a fixed-focus spectrograph, widely known as the Hilger Size C or E 3 instrument. In this instrument simple quartz lenses only were used, the camera lens being a separated doublet, and the photographic plate was mounted at a considerable angle to the axis of the camera to allow for the chromatic aberration of the lenses. The focal plane was, however, sufficiently flat for the whole spectrum from 2100 to 7000 A. to be obtained in good definition upon one plate which had the requisite slight curvature imposed upon it by the plate-holder. Rotation of the plane of polarization by the quartz was eliminated by the use of a Cornu prism and by adjustment of the thicknesses of the lenses. The inclination of plate proved to be no detriment, while the effective length of spectrum was considerably increased. Furthermore, as no alteration of adjustment was required in use, a scale of wave-lengths could be made, which, in its most convenient form, was photographed on to the spectrum negative itself, and this eliminated the laborious calibration of the instrument by the user. It

is not surprising that this type of spectrograph should play a part in popularizing spectrum analysis, nor that it should have been copied by other manufacturers.

While the principles of design outlined above have been adhered to there have been many changes in the details of design during the last ten years. As regards optical design, the camera lens has received considerable attention. The first definite improvement upon the original two-component lens was a single lens of which the surfaces were worked to aspherical curves, the working being controlled by the lens-testing interferometer. Besides an improvement of definition, there was a notable increase in speed, as this lens was capable of working at full aperture, while in the earlier one the definition depended upon the size of stop introduced between the components, and usually such a size of stop was selected as to effect a compromise between definition and speed. More recently the design has been elaborated so as to produce a really flat field, thus removing the restriction upon the thickness of photographic plate which could be used and eliminating any residual curvature either in the plate-holder or after removal. Such points may become of importance when, as frequently happens, the plate is to be submitted to examination on a microphotometer.

Turning now to the mechanical design, the early instrument was allied to its predecessors in having a wooden camera which supported the dark-slide and plate-holder. Various methods have been adopted for supporting the dark-slide on the base carrying the optical elements, but the most satisfactory design is one in which the construction is entirely of metal, thus providing the robustness required by industry. Minor points include the ease of removal and replacement of the plate-holder and facility of movement and positioning.

A spectrograph is, however, of no use by itself; it requires in addition a light source and, usually, also a condensing lens or other parts. Such an outfit tends to become fixed for any particular investigation or control, and it is not surprising, therefore, that many makers have introduced some form of optical bench or bar upon which this accessory gear is mounted and which is preferably rigidly coupled to the spectrograph, so that spectrograph and accessories are essentially one unit.

The instrument which has been discussed is well adapted to the investigation of non-ferrous metals and for much other general spectrum work. When, however, more complicated spectra, such as those of steels, elements of the platinum group, etc., are to be investigated considerably greater dispersion is required and in order to avoid uncomfortably large dimensions an instrument of the Littrow pattern is generally chosen. The optical arrangement is well known, but it may be pointed out that to pass from one region of the spectrum to another three separate adjustments are usually required. These are (*a*) rotation of the prism, (*b*) displacement of the prism carriage which also carries the lens, and (*c*) alteration of tilt of the plate-holder. Until recently these adjustments have been performed independently, and the change of spectrum range has involved the setting of three scales to predetermined values or to settings derived from calibration curves. To make the instrument more in keeping with industrial needs some of these adjustments have been combined by the makers and operation is performed from the plate-holder end at which

is situated also the light source, etc. In the case of one particular instrument, all adjustments are combined and are linked with a scale showing the range of spectrum which will appear on the plate at any setting. Such an instrument can be kept in operation all day, and the technique of spectrum analysis is now so advanced that check can be made of manufacturing operations at various stages.

The production of spectrum photographs, although a fundamental procedure, does not comprise the whole process of modern spectrum analysis. At the present time it seems as if the spectrographic equipment will in every case be linked with a microphotometer. This will suggest to many an instrument in which a photographic record is made on a greatly enlarged scale of some small portion of the spectrum negative. To perform analyses at a sufficient rate to control manufacture does not allow of a second photographic process and subsequent measurement. The instruments are, therefore, non-recording, and the sensitive element is a photo-cell of the rectifier type coupled to a robust galvanometer of short period. An enlarged image of the spectrum is thrown on to a screen provided with a slit behind which the photo-cell is placed, and mechanism is provided for traversing the line to be measured at a sufficiently slow speed for the minimum deflection to be accurately observed. A quantitative determination consists of such a measurement on two selected lines only in the spectrum—the one belonging to the element to be determined and the other to the main constituent. A very high speed of working has been claimed by Vincent and Sawyer[8], who used a Littrow spectrograph and a microphotometer for the routine analysis of steels.

In the spectrographs which have been discussed, the lenses are made only of the same material—crystalline quartz—as the prism. The considerable inclination of the photographic plate to the camera axis is determined by the relation of the chromatic aberration of the lens system to the dispersion of the prism. But, as the chromatic aberration is another manifestation of the dispersion of the material, it is to be expected that the inclination of plate will remain the same if some other material than quartz be used for both prism and lenses. This is, indeed, found to be the case, and hence with suitable design a glass system may be substituted for the quartz and the visible spectrum photographed with about three times the original dispersion, using the same instrument. Not only is this true of the 60° prism spectrograph, but also of the Littrow pattern with one 30° prism, and a model of the instrument with automatic adjustment is available in which both quartz and glass systems are mounted so that either can be used at will.

The advantage to be derived from this choice of optical system in the case just cited is increase of dispersion in a more limited range of spectrum. The principle is also of value when measurements are required near the limit of transparency of one system. In the case of a quartz spectrograph the limiting wave-length at the ultra-violet end depends upon the size of the instrument, particularly upon the size of prism. For an instrument of 20 cm. focal length for the *D* lines, working at, say, *F*/10, the spectrum can be photographed as far as 1850 A., which is also near the limit of transparency of air. For many purposes, however, it is of interest to pursue the investigation a further 300 or 400 A. into the Schumann region. For this it is

possible to exchange the prism and lenses of quartz for parts made from fluorite without change of adjustment of the instrument and then to place the whole spectrograph in a chamber which can be evacuated in order to obtain photographs.

A further instance of the same principle can be cited in the case of monochromators. It should, however, first be pointed out that a different aim underlies the design of these instruments from that which obtains in the case of spectrographs. For the latter, it is to obtain as long a range of spectrum as possible in good focus upon a flat plate, and as we have seen this can be achieved by placing the plate at a considerable tilt to the camera axis. For a monochromator it is desired that one line at a time shall be obtained of maximum strength and purity. The use of achromatic lenses in the ultra-violet is generally avoided on account of the inevitable secondary spectrum, while simple lenses can be set by a calibrated focus scale. But for a 60° prism monochromator this involves the setting of three scales as in the case of early quartz spectrographs. A small percentage of light of all wave-lengths is reflected from each optical surface, however well it may be polished, and thus a certain range of spectrum, centred upon the desired wave-length, passes through the exit slit of a monochromator. On this account it has become increasingly the practice in recent years to use double monochromators, but, as will be seen, this necessitates no less than six adjustments which, if performed separately, would form an intolerable burden upon the operator. Various means have been adopted by different designers to overcome this by coupling certain of the adjustments together, for example, the focusing of all four lenses. By adopting the Littrow form, however, the adjustment of prisms and lenses can be coupled as in the "automatic" spectrograph mentioned above, and this has been done in the Müller-Hilger double monochromator which is operated by one handle. For any one optical train the scale geared to this handle may be calibrated in wave-lengths, but where several trains are to be used alternatively this scale may be in terms of refractive index from which the position in the spectrum when using any one train may be deduced.

In connexion with the instruments we have been discussing, there are two points which deserve further mention—reflexion at the optical surfaces and absorption by the materials. The features generally to be desired in an instrument are large dispersion and large focal aperture. The former usually entails long focal length, which, if combined with the latter, also means large optical parts. Apart from the question of cost thus involved, we have already noticed that the limit of spectrum is set by the absorption of the optical material, and although this may be gradual it will be seen that a fairly definite limit is set even for prisms of moderate size. The path length is of course small near the prism apex, and when a large instrument is used to record near the limit it may well have a much smaller effective aperture than the apparent aperture. Multiplication of parts, particularly of prisms, will accentuate this. Reflexion losses have also been noticed, and it may be added that these are greater for oblique surfaces than for normal incidence which will again reduce the advantage to be derived from multiple-prism instruments.

There are cases in which mechanical design cannot be invoked to overcome the inconveniences of chromatic aberration. One which is of importance is the focusing

of interference bands upon the slit of a spectrograph, it being necessary that the bands of several lines should be in sharp focus at once, while the conditions of experiment render it desirable that the whole range of spectrum which can be photographed on one plate should be of good definition. The quartz-fluorite-sylvine quadruplet lens already mentioned was designed for use in this way with the Williams reflexion echelon. To obtain sufficient separation of the bands in the ultra-violet a lens of considerable focal length (1 m.) is required. The dimensions of the echelon, however, involve an aperture ratio which is not very high, and thus times of exposure make the repetition of photographs undesirable.

This instrument is customarily used in an evacuated chamber. The technique adopted with transmission echelons of inclining them in order to pass from single-order to double-order position is here inadmissible, as the reflected image is immediately lost. Whereas, however, the retardation with a transmission grating is the difference of the path in one plate and the corresponding path in air, with the reflexion grating it is twice the path in air corresponding to one plate. A variation of the density, or more exactly of the refractive index, of the air will therefore effect the required change of order, but it is necessary to hold the pressure of air constant throughout an exposure. An extension of this technique permits of the exact measurement of a wave-length *in vacuo*, thus eliminating uncertainties of temperature, pressure and humidity at the time of measurement.

The more usual interferometer for such measurements is the Fabry-Perot interferometer, customarily used in air and with such separation that focal lengths of 50 cm. are seldom exceeded for the focusing lens.

Considerable attention has been given in the last few years to redeterminations of the metre, or, alternatively, of the primary standard wave-length, i.e. that of the red cadmium line. The classical determinations by Michelson and by Benoit, Fabry and Perot were both made in France, and, while the agreement of their results is satisfactory, it was felt that redeterminations in other countries were desirable to establish that copies of the metre certified in Paris were still exact after transportation to their destinations. The Imperial Japanese Government, therefore, undertook such a redetermination using an exact copy of the apparatus employed by Benoit, Fabry and Perot. In this a series of comparisons is made between etalons of which one is of exactly one-half the length of the other, this exact ratio being shown by the production of superposition fringes in white light, and four such comparisons being necessary to pass from an etalon 1 m. in length to one 6·25 cm. long which can be directly evaluated in wave-lengths of light. Results published by Watanabe and his collaborators show satisfactory agreement with the earlier results, thus giving confidence in the Prototype metres which are issued. They also extended the wave-length measurement to a 5 m. geodetic base and modified the method so that a Jådrein wire 25 m. long could also be evaluated in wave-lengths[9].

Superposition fringes can also be obtained with etalons whose lengths are in any other exact ratio than one-half—say one-third or one-quarter—and in the corresponding redetermination made at the National Physical Laboratory[10], the comparisons have been reduced to two, while the length of etalon directly evaluated was

8·33 cm. The method of holding the etalon plates to the separators by spring pressure was superseded by wringing the plates on to the polished ends of invar tubes, while the intermediate standard was designed as an end-gauge for comparison with the longest etalon by interference. The ultimate comparison of an interference standard with a line standard continues to be the weakest link in the chain, although it has been strengthened from the somewhat involved operation adopted in the Fabry-Perot method.

Great attention was paid in the design to elimination of uncertainties in air conditions such as humidity and CO_2 content, and determinations *in vacuo* have been made with this apparatus. Determinations of importance to spectroscopists which can be made with this apparatus are those of the refractive index of air for various wave-lengths. The previous best were those made by Meggers and Peters[11], and it has long been felt that measurements by a different method are required.

Another instrument which has been suggested for this purpose is the Rayleigh Interference Refractometer. In the form in which it is customarily used, it is not well adapted, as fringes in white light are employed and settings are made on the achromatic bands at the centre of the system. Individual bright spectrum lines can be used, but the luminosity does not in most cases allow of projection of the fringes upon the slit of a spectrograph. In a modification due to Williams[12] light is taken from adjacent parts of the collimator objective which allows of the observation of bands with slit widths nearly three times that previously permitted. Applying the same principle at the telescope end also, the resultant brightness does permit of projection of the bands. No results appear, however, to have been published using this device.

Various aspects of chromatic aberration as it affects the ultra-violet in particular have been discussed in what has gone before. When attention is turned to the infra-red end of the spectrum it is found that the solution is rather sought in its elimination by using mirrors than in any form of lens design. This may be in part attributed to the fragile nature of the substances useful in this region and partly to the fortunate circumstance that it is not difficult to find substances with high reflecting power for these wave-lengths, which has certainly not been the case for the ultra-violet until quite recently, and even now the best reflectors leave something to be desired in the latter region. The use of mirrors leads to a shape of instrument unlike that embodying lenses, but again the substitution of one prism by another can generally be effected with ease. High resolution when required is generally obtained with a ruled diffraction grating, and to avoid the confusion of overlapping others a prism spectrometer is often coupled to the grating instrument, thus leading to the combined prism-grating instrument first described by Sleator and Imes, of which several modifications have also been noted in the literature.

One of the major problems in this region is, however, to make instruments of high light-gathering power, and in this connexion one has been described by Randall[13] in which part of a concave mirror 24 in. in diameter and of focal length 36 in. is used. This was sufficient to fill a grating with ruled surface 10 × 22 in., as large as can be handled at the University of Michigan which specializes in work in the

infra-red. Naturally both large mirrors and large gratings are costly, but it seems likely that apparatus on this scale, or even larger, will be required to deal with the region beyond that which can be investigated by the prism materials already to hand. The problem is also complicated by the source of radiation, for while the intensity in the long-wave region is increased by raising the temperature, this is accompanied by a much more rapid rise in the shorter wave region, and this latter energy must be in some way dissipated in order to prevent interference in the longer wave region.

Another aspect in which the infra-red region appears to differ from the others is that the detector—usually a thermopile—is a direct measure of intensity, so that a spectrometer for this region is also a spectro-photometer, and results are always, to some extent at least, quantitative.

In the visible spectrum, the polarization photometer continues to be used, its application to industry being more appreciated. Extinction coefficient being the quantity of ultimate interest, it is customary to provide a scale of density, i.e. $\log_{10} I_0/I$, connected to the moving prism. When the control beam only is cut down the density is proportional to $\cos^2\theta$, θ being the angle through which the prism is rotated, and while the scale is very open for low densities, it is cramped for high ones. The introduction of another prism into the beam where the sample is placed, this prism being set with its azimuth at right angles to that of the polarizer in the control beam, reduces the angular length of scale from 90° to 45°, while the law obeyed is now $\tan^2\theta$ instead of $\cos^2\theta$. The new scale is much more open relatively at the high densities, while the setting is more critical than with the earlier type. Furthermore, any polarization which may be present in the source is taken care of.

In this instrument, however, as in the cases of others already mentioned, there is a certain inevitable scattering of light which may become apparent when quite high densities are to be measured. As in the case of a monochromator where this has been overcome by a second dispersing system, so in this case it is overcome by a second polarizing system, and such an arrangement has been described by Dowell[14]. A fixed polarizing prism with azimuth parallel to that of the first follows the rotating prism in each beam, and the law now obeyed by the instrument is $\tan^4\theta$ instead of $\tan^2\theta$. The angular length of scale is still 45° and the scale is still more open relatively at the high density end. For precision work at high densities this instrument appears to be the best yet put forward.

Some experience in the use of such instruments seems to be essential for obtaining most accurate results, and one of the greatest limitations to the use of instruments of this class appears to be fatigue in the operator. This is likely to become evident after a lapse of time controlled by several personal factors, and readings made under conditions of fatigue are not comparable with those made under good conditions.

It is probably due to considerations such as these that there has been an impetus towards the design of recording instruments in which the sensitive element is some form of photo-electric cell. Such instruments appear to reproduce their results with satisfactory precision, and of course are available for work throughout the day. In

spite, however, of the great use to which such cells have been placed and the consequent study of their properties under many conditions, there still seems to be a liability to fatigue effects and the working conditions need to be definitely laid down and adhered to if accurate results are to be obtained. An allied instrument which seems likely to have a large field of application is a photo-electric absorption photometer in which the region is selected by filter only and which would replace the visual instrument known as a "Duboscq" colorimeter. With the latter it is sometimes required to match one solution with another of slightly different tint. This would lead to various observers getting differing results, but the objective instrument would yield the same reading with all operators.

Photo-electric instruments can also be used in the ultra-violet, one in which the principles of design seem soundest being due to von Halban[15]. The density measurements are made with a sector, two cells being used in opposition to find the balance point of the two beams. No assumptions need to be made concerning the linearity of response to intensity, and the avoidance of such assumptions appears to be at the base of sound design.

For years before photo-electric cells were practicable instruments, photometric measurements were made involving the use of the photographic plate. The best-known photometer of this class is the Hilger sector photometer which has become a standard item in the outfits of several manufacturers and was adopted by the Bureau of Standards[16] as the recommended instrument for absorption measurements in the ultra-violet. It embodies two sectors, one in each beam, which are driven at the same speed by an electric motor. The sector in the beam in which the substance is placed is usually invariable, while that in the control beam can be set to a predetermined value of opening, thus giving a definite ratio of times of exposure to the two beams.

There are, however, a number of inconvenient features. The motor must be stopped in order to change the sector opening and allowed to attain its normal speed before making an exposure, the moving parts may introduce some vibration, the sectors at maximum aperture cannot transmit more than 50 per cent of the incident light, and the exposure is intermittent. To remove these, a form of photometer was introduced by Judd Lewis which did not involve moving sectors, and the time lost between successive exposures was greatly reduced. The form of shutter incorporated in this instrument was of elegant design, but proved to be costly to manufacture. The Spekker photometer incorporates a simple shutter which has been proved to be accurate, and the times of exposure necessary have been considerably reduced.

Several devices have been put forward for obtaining fine bands in absorption using a continuous spectrum. A recent one, due to Holiday[17], appears worthy of mention. In it the spectrograph plate-holder is moved by a logarithmic cam so that exposures decrease logarithmically from a datum line. It appears to be well adapted for use with instruments having a flat field, so that no ambiguity arises from any curvature to which the plate has been subjected.

REFERENCES

(1) Williams, W. E. *Z. Phys.* **53**, 542 (1929).
(2) Jackson, D. A. and Kuhn, H. *Proc. roy. Soc.* A, **148**, 335 (1935).
(3) Williams, W. E. *Proc. phys. Soc.* **45**, 699 (1932).
(4) Perry, J. W. *J. sci. Instrum.* **9**, 116 (1932).
(5) Stockbarger, D. C. *Rev. sci. Instrum.* **7**, 133 (1936).
(6) Smith, C. F. *J. sci. Instrum.* **10**, 272 (1933).
(7) Bricout, P. *Rev. Opt. (théor. instrum.)*, **10**, 345 (1931).
(8) Vincent, H. B. and Sawyer, R. A. *J. appl. Phys.* **8**, 163 (1937).
(9) Watanabe, N. and Imaizumi, M. *Proc. Imp. Acad. Japan*, **5**, 233 (1929).
(10) Sears, J. E., Jr. and Barrell, H. *Philos. Trans.* A, **231**, 75 (1932).
(11) Meggers, W. F. and Peters, C. G. *Bull. U.S. Bur. Stand.* **14**, 697 (1918).
(12) Williams, W. E. *Proc. phys. Soc.* **44**, 451 (1932).
(13) Randall, H. M. *Rev. sci. Instrum.* **3**, 196 (1932).
(14) Dowell, J. H. *J. sci. Instrum.* **8**, 382 (1931).
(15) von Halban, H. and Ebert, L. *Z. phys. Chem.* **112**, 321 (1924).
(16) Sci. Paper No. 440, *Bull. U.S. Bur. Stand.* **18**, 121 (1922).
(17) Holiday, E. R. *J. sci. Instrum.* **14**, 166 (1937).

SPECTROSCOPY

Edited by W. JEVONS, D.Sc., Ph.D.

§ 1. SPECTROSCOPIC METHODS OF CHEMICAL ANALYSIS

By W. R. S. GARTON, B.Sc., A.R.C.S., Imperial College of Science and Technology

Introduction. The possibility of basing a system of qualitative chemical analysis on spectroscopic observations appeared from the work of Bunsen and Kirchhoff and others, which led to the discovery that a chemical element emits a characteristic line spectrum, even when other elements are present. Thus once the spectrum lines of an element were known they could be used in order to determine its presence in a mixture. The method at once proved its power in leading to the discovery of a number of new elements, e.g. Rb and Cs by Bunsen in 1860, but the interest of chemists in spectrum analysis appears soon to have waned, and the development and systemization of the subject was left to a few isolated workers.

The feasibility of a system of quantitative spectrum analysis is suggested by the observation that when the percentage of an element in a mixture is varied there is a corresponding variation in the intensity of its spectrum lines, as a rule in the same sense. Several early attempts were made in this direction, e.g. Lockyer's method of "long and short lines", based on his observation that in the spectrum of a spark between electrodes of a metallic alloy the lengths of the lines depended on the concentrations of the elements concerned. The establishment of methods of wider application had, however, to await a better understanding of the processes involved in the excitation of atoms, and it is within the last decade that appreciable progress has been made. During this time a large amount of work has been done both on the qualitative and on the quantitative side, and more or less definite systems of analysis have been devised, largely through the labours of Gerlach and his collaborators, of Lundegårdh and of Twyman and his associates, amongst others, and a considerable amount of literature has become available, notably the volumes by Gerlach and collaborators[1], Lundegårdh's account of his researches[2], various publications by English workers[3] and an excellent recent work by Swings[4].

Methods of excitation. Various forms of flame, arc and spark sources have been devised and employed by different workers for the purposes of analysis. The most important work performed with flame sources is that of Lundegårdh, who used the air-acetylene flame in an extensive series of quantitative investigations in biological and soil chemistry. In quantitative work the control and uniform maintenance of the excitation conditions is essential. Accordingly, in the apparatus used by Lundegårdh, the gas and air pressures are controlled, and the substance to be analysed is introduced into the source, either as a powder or in the state of solution, by an atomizing arrangement, the apparatus being carefully designed in every detail to ensure reproducibility of results.

The low-tension arc in air is a particularly useful source in qualitative analysis, and has been used, for example, by Goldschmidt[5], by Breckpot[6] and by Lewis[7]. The arc may also be of use in quantitative work in cases where the samples are in the form of conducting rods, as, for example, in alloy testing; but difficulties appear when auxiliary electrodes have to be employed. One of these arises from the differential volatilization of the elementary constituents of a mixture when these have widely different boiling-points. This effect has been studied at length by Mannkopf and Peters[8], and by Breckpot[6], who found that it could be largely avoided by the use of graphite electrodes, the sample being placed in a cavity in the cathode, and the current being kept low. In qualitative analysis the effect has been turned to good use, as it tends to enhance the lines due to volatile elements present in traces; in this case the current is raised and the substance is placed on the anode. In this way Gerlach increased by a factor of one hundred the qualitative sensitivity in the detection of traces of lead in gold. Mannkopf and Peters advise the use of that portion of the discharge in the immediate neighbourhood of the cathode, as they found there a considerable intensification of the lines of elements present in small concentrations as well as a diminished band intensity. Hitchen[9] has tried to avoid the unsteady running of the arc which sometimes results from the formation of oxide by directing a small reducing flame upon it.

The most extensively used method of excitation is that provided by the condensed spark discharge from an induction coil or transformer, employed in particular by Gerlach in his earlier work. In quantitative work, since the results obtained are sensitive to variations in the frequency of interruption, care must be taken that any form of mechanical interrupting device used shall run steadily. The use of an inductance in the secondary circuit facilitates adjustment of the excitation conditions, while serving to suppress band spectra. The best quantitative results are probably obtained when the substance to be analysed is in the state of solution; and various forms of spark-gap designed to provide uniform and reproducible conditions in such cases have been proposed, e.g. by Twyman and Hitchen[10], by Burke[11] and by Jolibois and Bossuet[12].

The intermittent arc has been used, again principally by Gerlach. In this arrangement the arc is produced by the separation of two electrodes, one being attached to a cam, so that about five contacts per second are made. The method is said to be satisfactory in the analysis of easily fusible alloys, effects of fractional distillation being avoided owing to the cooling which occurs during interruption. Gerlach and Gerlach have described a high-frequency Tesla arrangement which is particularly suitable for the analysis of non-conducting samples such as organic materials and pathological specimens, e.g. bone splinters, and makes possible the avoidance of preliminary chemical treatment of the specimen.

Qualitative analysis. The procedure adopted in order to analyse qualitatively a given sample of material by spectroscopic methods is relatively simple. The substance is excited in one of the ways mentioned, and the spectroscope or spectrograph is used to identify the elements present by the wave-lengths of the lines observed, determined by means of a known comparison spectrum, usually of the iron arc.

The origins of the observed lines are determined from tables, of which many arranged for use in qualitative analysis are now available[13]. The technique can be simplified in some cases by the use of a calibrated wave-length spectroscope or spectrograph, and in routine work an operator with a little practice becomes able to recognize many lines at sight.

The Hilger Spekker Steeloscope[14], designed primarily for the routine testing of steels, forms a good example of a specialized spectroscope for industrial analyses. The eyepiece of this instrument can be moved in the direction of dispersion, and certain points on the slide carrying the eyepiece are marked with the symbols of elements having strong lines in those positions, this calibration having been performed by the makers. The apparatus can also be used in performing rough quantitative analyses of samples of fairly constant composition, as in the case of steels where there is a great preponderance of one element, viz. Fe, and it has found considerable application in industrial laboratories.

Although the visual examination of the spectrum of a sample is often sufficient for the purpose in hand, photographic methods are more comprehensive, partly because the ultra-violet region of the spectrum is made available, many elements having their strong lines in this region, and also because faint lines are recorded, and the corresponding presence of small traces of elements recognized. The labour involved in interpolation can sometimes be reduced by using a series of comparison exposures, if the presence of particular elements is suspected, or by the use of a spectrograph of the type in which a wave-length scale is impressed on the plate during exposure. Another valuable labour-saving device is to be found in Ryde and Jenkins *R.U. powder*[15], which consists of a mixture of small quantities of about fifty elements diluted with a base of alkaline earth oxides, the quantities of the various elements being adjusted so that only their most sensitive spectrum lines* appear. The spectrum of this powder is photographed alongside that of the test sample, and the composition of the latter is found by inspection. These methods make possible the detection of all the metallic elements and some non-metals, e.g. B, As, P; but with several non-metals such as S, with permanent gases and with acid radicles they are unable to deal.

It is of importance to know the precision of which these qualitative methods are capable, and although the figures vary from one element to another as well as with the means of excitation adopted, the following points are of interest. If the relative concentration of an element in a mixture is small, not all the lines of that element will appear in the spectrum of the mixture, the weaker lines being lost in the background of scattered light and continuum; only the most persistent lines will be observed. These are, of course, the last lines to disappear as the concentration of the element tends to zero, and were called by de Gramont the *raies ultimes*. As a rule the *raie ultime* of a given element is the intrinsically most intense line, or the corresponding multiplet, in the complete arc spectrum, although certain exceptions occur, as for instance when self-absorption takes place. Again, in certain cases the most persistent lines may belong to the ionized atom. The *raies ultimes* are, of

* Their *raies ultimes* (R.U.); see below.

course, the lines which are generally of greatest use in qualitative analysis, but care is sometimes necessary in inferring the absence of a particular element from the absence of the catalogued *raie ultime*. The limiting concentration at which an element is detectable in a mixture is set by the disappearance of the most persistent lines into the background of diffuse light, etc.; the occurrence of a high background intensity will reduce the analytical sensitivity, and, if particularly bad, as in the neighbourhood of strong bands, may render a certain spectral region useless. The figure for this lower limit is thus subject to a number of factors, but lies as a rule between 10^{-3} and 10^{-5} atomic per cent, according to the element considered and the method of excitation employed. A possible source of error results from the occurrence of coincidences, which may make it difficult to decide whether a particular line is a weak line of a major constituent or a strong line of a minor.

Quantitative analysis. The task set in quantitative analysis is to determine how many out of one hundred atoms of a specimen belong to elements A, B, etc., by an examination of the intensities of their spectrum lines. As yet there has been no comprehensive system of this kind developed, and each problem has to be studied individually. This largely results from the fact that the absolute intensity of a spectrum line is no measure of the concentration of the corresponding element, since it is a function of the type of, and conditions in, the source used, and is also dependent on the nature of the spectroscopic apparatus employed. The problem is much simplified if it is sought to establish the relative proportions of the elements present by a consideration of the intensity ratios of their lines. Even so the intensity ratio of two lines, one due to each of two elements in a mixture, is very much affected by the excitation conditions, and hence by the simultaneous presence of other elements in the source. This is particularly the case if one of the comparison lines refers to a neutral and the other to an ionized atom. Accordingly, except in the simplest cases, one of the first requirements is that some means should be available of controlling and maintaining uniform the excitation conditions, and it is in this direction that most work has had to be done.

The lines used in quantitative work should have high *concentrational sensitivity*, this quantity being conveniently defined as dI/dc, where I is the intensity of a line and c the concentration of the corresponding element. This condition is very often not fulfilled by the most persistent lines when the concentration c is appreciable, and it is generally advisable to avoid lines involving transitions to ground states on this account.

All the methods that have been devised depend on interpolation between or matching with a series of standard samples, the *method of comparison spectra* being the most obvious procedure of this kind. This method consists in impressing on a plate a series of spectra of standard mixtures together with the spectrum of the test sample, and finding by inspection which of the former most nearly matches the latter. Interpolation can, of course, be performed, and the method can be improved and extended by the use of some form of apparatus for the measurement of photographic densities, and can be used successfully where a large number of analyses of mixtures have to be made, in which only one elementary constituent is

subject to appreciable variations. The introduction of large variations in the percentage of a third element present makes the process too cumbersome. The method has been modified by Lundegårdh, who used a photoelectric cell at the focus of a spectrograph to compare intensities directly. It is essential that the spectra should be photographed under the same conditions, and a neat way of ensuring this is found in Barratt's "twin-spark" method, in which the spectra of two sparks in series are observed or photographed simultaneously.

The *method of homologous pairs*, principally developed by Gerlach, provides a means of establishing tables which can be used in subsequent analyses with other instruments, so that the trouble of preparing standard samples and photographing comparison spectra every time is avoided. Obviously the first requirement is some criterion by which the excitation conditions can be reproduced, and Gerlach has secured this by making use of the variation in the intensity ratio of two lines of one of the major elementary constituents of the mixture, when one of these belongs to the neutral and the other to the ionized atom. Thus, for example, if the spectrum of a mixture of two elements P and A containing x per cent of A is photographed it is often possible to select two P lines λ_P', λ_P'' occurring close together on the plate and of equal intensity, this equality, moreover, being greatly affected by variations in the conditions in the source. In order to reproduce subsequently these same conditions all that is necessary is to adjust matters so that these two lines are of equal intensity. Once this so-called "fixation pair" has been decided on, other pairs of lines λ_P, λ_A are selected in the P and A spectra, the components of each pair being fairly close and having equal intensities for a particular value of the relative concentration x. These latter lines constitute what are termed *homologous pairs*, and are chosen so that the intensity equality is as little affected as possible by changes in the excitation conditions. Homologous pairs having been found for a series of values of x, the results are tabulated and may be used in future analyses. Although theoretical objections to the validity of the fixation-pair criterion readily suggest themselves, it seems to work well enough in practice, and was fully tested by Gerlach. The method can be made more accurate by the provision of interpolation tables, obtained either visually or by the use of a photometer, and in this way Gerlach's tables[1] have been extended. The use of internal standards is also often of use, as for instance when no homologous pairs for elements P and A are known, although tables for P and B have been drawn up. In this case a measured quantity of B can be added to the sample, and the latter tables used.

For the method of homologous pairs there must be a sufficiency of suitable comparison lines. Often one of the elements P or A is lacking in this respect; Gerlach has devised what he calls *substitution methods* to deal with such cases. The essence of the procedure consists in artificially increasing the number of P or A lines available, and is illustrated by the following example. Suppose it is desired to establish analytical tables for associations P + x per cent A of two elements, and that the element P presents no lines suitable for homologous pairs. The spectrum of the mixture is photographed and alongside it is impressed the spectrum of another element S, called the *substituted spectrum*, the latter being excited under definite

conditions specified by a fixation pair. The two spectra are now related or "coupled" by choosing two lines λ_P', λ_S' belonging to P and S and adjusting the relative exposure times so that these lines produce equal plate blackenings. It will now probably be possible to find line pairs λ_A, λ_S of equal apparent intensity at concentration x of A in the mixture, i.e. an homologous pair at this concentration, and the process having been repeated for various values of x, tables can be drawn up and used in future work, the conditions under which the spectra have been examined being defined. The method can also be usefully applied to extend existing tables of homologous pairs, by lowering the limit of quantitative detection, the spectrum of P itself often then acting as the substituted spectrum. Thus the obstacle preventing extension of the tables for the mixture of elements P and A to lower relative concentrations might be a deficiency of weak P lines, and by using a substituted spectrum this deficiency may be overcome.

Gerlach decided, as the result of a large number of experiments, that the concentrational sensitivity is the same for all lines of any one element, except those showing self-reversal, and is independent of the simultaneous presence of other elements. Methods of analysis have been devised which depend on these facts; the rather complicated procedure is fully described in the work cited, but cannot be summarized in a few words.

Wendt and Heun[16] vary the inductance in a condensed spark circuit so as to equalize the intensities of two chosen lines for various relative concentrations; having performed a preliminary calibration, they are able to use the variable inductance to measure the concentration in a test sample. The method is applicable only in cases where no appreciable variation occurs in the amount of a third element present.

Photometric methods. Although the visual examination of spectrograms is sufficient for many purposes, the accuracy of the quantitative methods described can be improved by the use of some form of instrument for measuring relative intensities. The most accurate procedure is to use some form of microphotometer for this purpose, as has been done by Lundegårdh, but there are several simpler devices available. One of these[17] consists of a neutral wedge of glass or quartz, which is placed before the spectrograph slit so that lines of different intensities appear of different lengths on the plate, and the difference in length of lines due to two elements can be correlated with the percentage of one of them present. The same end is achieved by the use of the logarithmic sector, first applied to quantitative analysis by Scheibe and Neuhausser[18]. This consists of a disc, having its circumference cut in such a shape that, when it is rotated in front of the slit, the effective exposure time varies logarithmically along the length of the slit, and the length of a line on the plate is a logarithmic function of the intensity. The method has been developed by Twyman and Simeon[19] amongst others, and the photometric aspect has been discussed by Martin and collaborators[20]. In the use of both the wedge and the log. sector, the accuracy is limited by the difficulty of fixing the exact end-points of the lines. In a recent development of the wedge method by Foster and Horton[21] the location of the end-points is obviated.

The accuracy obtainable by the various methods of quantitative analysis described is, of course, a very variable quantity, but general figures may be given. Thus it is possible, as a rule, to make estimations by the method of comparison spectra or by that of homologous pairs with a precision of 5 to 10 per cent, and this may be improved to 4 or 5 per cent by the use of one of the photometric devices; indeed Lundegårdh claims to have achieved an accuracy between 1 and 3 per cent. These latter figures suggest that, in order that the methods may be developed in the direction of improved accuracy, some improvement in the uniformity of photographic plates will become necessary, as it is generally considered that irregularities in the plates can introduce errors of the order of 2 per cent.

Summary. Spectroscopic methods of analysis supplement rather than compete with the ordinary chemical methods. They effect a saving of much time and labour in those cases of routine work where very great accuracy is not required, and have the advantage that the amount of material required may be very small, while all the elements present, subject to the limitations already referred to, may be detected and estimated simultaneously. The uniqueness and real power of the methods are manifest in those cases of micro-analysis in which the quantities of elements involved are so minute as to be undetectable chemically. Even where a complete accurate analysis of a large bulk of material is required, and chemical methods are indispensable, a preliminary spectroscopic examination is often of use, as it will give an indication of nearly all the elements present, and roughly their proportions. These points explain the increasing extent to which the methods are being applied in (to mention a few fields) metallurgical work, problems of biological, pathological and agricultural chemistry, and mineralogy and geo-chemistry. The chief restriction of the methods is inherent in their empirical nature. The ideal system would probably be one for which an accurate physical theory could be developed, perhaps based on absolute intensities; Fowler[22], in a recent address, has put forward some interesting speculations in this connexion.

References

(1) Gerlach, W. and Schweitzer, E. *Foundations and Methods of Chemical Analysis by the Emission Spectrum* (Hilger, 1931), English translation of *Die Chemische Emissionsspektralanalyse*.

Gerlach, W. and Gerlach, W. *Clinical and Pathological Applications of Spectrum Analysis* (Hilger, 1934), English translation of Part 2 of above.

[Each of the above volumes contains an extensive bibliography.]

Gerlach, W. and Riedl, E. *Die Chemische Emissionsspektralanalyse*, Part 3 (Voss, Leipzig, 1936). [Consists mainly of tables for qualitative analysis.]

[Gerlach and collaborators have contributed also a considerable number of papers to various German chemical journals during the past few years.]

(2) Lundegårdh, H. *Die Quantitative Spektralanalyse der Elemente* (Fischer, Jena, 1934).

(3) Twyman, F. and collaborators. *The Practice of Spectrum Analysis with Hilger Instruments* (Hilger, 1935).

Smith, D. M. *Metallurgical Analysis by the Spectrograph* and *Bibliography of Literature on Spectrum Analysis*. British Non-Ferrous Metals Assoc. (1933 and 1935).

(4) Swings, P. *La Spectroscopie Appliquée* (Hermann, Paris, 1935).

(5) GOLDSCHMIDT, V. M. and others. *Nachr. Ges. Wiss. Göttingen*, various numbers since 1931.
(6) BRECKPOT, R. and others. *Ann. Soc. Sci. Bruxelles*, various numbers since 1933.
(7) LEWIS, S. J. *Spectroscopy in Science and Industry* (Blackie, 1933).
(8) MANNKOPF, R. and PETERS, C. *Z. Phys.* **70**, 444 (1931).
(9) HITCHEN, C. S. *Tech. Publ. Amer. Inst. Min. Met. Engrs*, No. 494 (1933).
(10) TWYMAN, F. and HITCHEN, C. S. *Proc. roy. Soc.* A, **133**, 72 (1931).
(11) BURKE, S. A. *Trans. Faraday Soc.* **33**, 309 (1937).
(12) JOLIBOIS, P. and BOSSUET, R. *C.R. Acad. Sci., Paris*, **204**, 1189 (1937).
(13) LÖWE, F. *Atlas der Analysen-Linien der Wichtigsten Elemente* (Steinkopf, Dresden and Leipzig, 1936).
TWYMAN, F. and SMITH, D. M. *Wave-length Tables for Spectrum Analysis* (Hilger, 1931). [Contains an account of the researches of de Gramont and other early workers.]
(14) Hilger pamphlet No. 243. *The Spekker Steeloscope.*
(15) RYDE, J. W. and JENKINS, H. G. *Sensitive Arc Lines of 50 Elements, including Notes on the use of the R.U. Powder in Spectroscopic Analysis* (Hilger, 1929).
(16) WENDT, A. and HEUN, H. *Z. Phys.* **93**, 92 (1934).
(17) NICHOLSON, J. W. and MERTON, T. R. *Philos. Trans.* A, **216**, 459 (1916).
(18) SCHEIBE, G. and NEUHAUSSER, A. *Z. angew. Chem.* **41**, 1218 (1928).
(19) TWYMAN, F. and SIMEON, F. *Trans. Opt. Soc.* **31**, 169 (1929–30).
(20) MARTIN, L. C., BURKE, S. A. and KNOWLES, E. G. *Trans. Faraday Soc.* **31**, 495 (1935).
(21) FOSTER, J. S. and HORTON, C. A. *Proc. roy. Soc.* B, **123**, 422 (1937).
(22) FOWLER, A. *J. sci. Instrum.* **14**, 153 (1937).

§ 2. ASTRONOMICAL SPECTROSCOPY

By A. HUNTER, Ph.D., F.R.A.S., Imperial College of Science and Technology

Interstellar matter. Investigation of the composition and distribution of interstellar matter may be approached in two ways. One—the earlier—starts from observational facts and attempts to build up from these a theory of the structure of the interstellar cloud. It is in this way, from the appearance and behaviour of the detached lines in stellar spectra, that most of our knowledge of interstellar matter has been derived. The other starts by assuming a cloud of matter of plausible chemical composition, and predicts the observable effects of such a cloud. From its very nature, this attitude cannot justify any assertions regarding the physical facts of the universe, but it may nevertheless prove very fruitful in directing observation into the most useful channels. Recent progress along both lines of approach justifies a survey of the present situation.

Over 30 years ago Hartmann first discovered that the *K*-line in the spectrum of the binary δ Orionis was "stationary", i.e. that it did not share the Doppler displacement, due to orbital motion, common to all the other lines. The inference of an interstellar origin was made at once. Since then the list of interstellar lines has been enlarged: it now includes the *H* and *K* lines of Ca^+, the *D* lines of Na, and, more recently[1], lines of unknown origin at 5780·4, 5796·9, 6283·9 and 6613·9 A. Observational study of the Ca^+ lines as they occur in various stellar spectra suggests that the absorption occurs in a stationary cloud of calcium ions distributed fairly uniformly throughout the galaxy. Theoretical calculation of the density, temperature and state of ionization of the cloud on this assumption predicts for the interstellar ines the right absolute intensity, the right galactic rotation effect and the right

behaviour along the stellar sequence, as regards blending with stellar lines. The theory also accounts satisfactorily for the absence of neutral Ca lines, and for the presence of the *D* lines in spite of the low ionization potential of Na. An exaggerated idea of the uniformity and tranquillity of the cloud, however, has recently had to be removed[2]: results for individual stars show anomalies which must be attributed to local irregularities[3]. The four lines of unknown origin present a problem whose solution will be awaited with interest. The visible ultimate lines of all atoms and ions which are likely to be abundant in interstellar space are now so well known that when an appeal to them proves fruitless a molecular origin is suggested[4]. The diffuse nature of the lines supports this view, and indeed coincidences with certain CO_2 vibration-rotation band-heads are reported[5] for the two lines in the red. A search for the "Venus bands" of CO_2 at 7820, 7883 A., which should be stronger than the ones thus tentatively identified, will decide whether or not this result is a chance coincidence.

From the theoretical point of view we start with a cloud of reasonable composition and try to explain the observed facts. The atomic abundances found in the stars, in the sun, in the planets, on the earth, in the gaseous nebulae—in short, in all the matter of the universe, show a remarkable similarity. Apparent divergencies are with few exceptions due merely to different physical conditions: temperature, pressure, density, degree of ionization, surrounding radiation field, etc. Starting, then, with a cloud of the general cosmic composition, we find many elements unobservable either because their ultimate lines are out of reach, or because they have such low ionization potentials as to be present only as ions with ultimate lines again out of reach. The probable preponderance of hydrogen, however, should have interesting consequences[6]. Its presence should, as in the gaseous nebulae on Zanstra's luminosity theory, gradually eliminate all radiation capable of ionizing it. Thus, except close to the hotter stars, no photons of energy greater than 13·5 v. should exist in the cloud. This in turn suggests that many of the common atoms will be neutral and capable of forming stable molecules.

Hence from both observational and theoretical considerations we are led to suspect the presence of molecules in interstellar space. It may even be that the "dark nebulae" visible in certain regions of the sky are simply multi-molecular aggregations in the general interstellar cloud.

Bright solar eruptions and radio fading. Among astronomical instruments of recent invention none has proved more productive of useful and suggestive results than the spectrohelioscope, a device which is scarcely more than 10 years old. To its precursor, the spectroheliograph, now 45 years old, we owe much of our present knowledge of the solar atmosphere. The spectroheliograph photographs the sun in the monochromatic light emitted by a single element (the red hydrogen line or the ultra-violet lines of ionized calcium are often used), and thus gives an idea of the distribution of the emitting particles over the disc. Comparison between spectroheliograms and direct photographs has led directly to the discovery of important relationships between such features of the sun's atmosphere as the flocculi, faculae, spots and prominences.

The spectrohelioscope, as its name implies, is a similar instrument adapted to visual observation. Though never likely to supplant the spectroheliograph, not the least of whose functions is to provide a permanent day-by-day record of the constitution of the sun's atmosphere at different points on the disc, the newer instrument is invaluable for studying rapidly changing solar phenomena in monochromatic light. Spectrohelioscopes are now in use all over the world, and international co-operation between the various observatories in which they have been installed has quite recently made continuous observation of the sun (given good weather) a realizable project: only a few half-hourly or hourly gaps in the twenty-four hours now remain to be filled[7].

Chief among the transient phenomena on which attention is now concentrated are *chromospheric eruptions*, in which the brightness of localized parts of the disc, as observed in monochromatic light, increases enormously within the space of a few minutes. The typical eruption reaches its maximum brightness, which may equal or even exceed that of the continuous spectrum adjacent to the line in which it is observed, in 5 or 10 min., and gradually fades completely in another half-hour. It seems to occur in a facular region just prior to the formation of a spot, and is of sunspot dimensions. It is never observable directly, in an ordinary telescope, but must always be observed in some spectroscopic way, either with a spectrohelioscope, or more simply by means of an ordinary solar spectroscope with a wide slit. During 1936, a year of rapidly increasing solar activity, it is calculated that, allowing for poor weather conditions, an average of four eruptions must have occurred per diem.

Recent observations have given rise to suspicions that these eruptions, like sunspots, have terrestrial consequences[8,9]. Periods of sudden attenuation in short-wave (about 20 m.) radio transmission, affecting the daylight side of the earth and lasting about half-an-hour each, appear to follow nearly all bright eruptions at intervals varying from 4 to 17 min. The material available points unmistakably to a causal connexion between eruption and fading, though difficulties due to the evanescent nature of both phenomena make it impossible to fix the time-interval between the occurrences exactly: it is of the order of 10 min. It is significant that no case is satisfactorily recorded in which a fade-out has preceded an eruption obviously connected with it. Also associated with these chromospheric eruptions are terrestrial magnetic disturbances. These take the form of sudden slight variations in the magnetic elements which may occur a few minutes after a bright eruption has been noticed. These changes should not be confused with the magnetic storms which occur some 30 hr. after the meridian passage of their associated sunspots: these latter disturbances are of a different order of magnitude both in intensity and in duration.

With the evidence that has now accumulated[9], and with the knowledge that the sun is 8·3 light-min. distant from the earth, it becomes an irresistible temptation to ascribe both radio and magnetic phenomena to electromagnetic radiation emitted from a solar eruption and disturbing the earth's ionosphere some minutes later. The fact that the eruptions effective in producing radio fading favour no

particular heliographic longitude supports this view. Unlike sunspots, which are supposed to cause magnetic storms by a radial corpuscular agency, and are therefore not terrestrially effective unless they are near the central meridian, the eruptions on this hypothesis emit radiant energy which reaches the earth regardless of its exact place of origin on the disc. Nevertheless, the theory is not without its apparent inconsistencies. Why, for example, does not *every* eruption produce high-frequency radio fading? Conversely, if sudden short-wave attenuation is to be regarded as caused solely by solar disturbances, why are not *all* fade-outs preceded by chromospheric eruptions? And if the hypothetical radiation is to be regarded as Planck radiation from a very hot area on the sun's surface, why are not the outbursts—so conspicuous in H or Ca^+ light—also visible on direct photographs in white light, or at least on monochromatic photographs at some wave-length in the continuum? It may be, of course, that these questions do not accurately reflect the facts. In any case, answers to them must await further observations during the period of increasing solar activity predicted for the next two years. On the solar side, good weather and more observing stations with spectroheliosopes are required. On the radio side, a programme of direct and continuous sounding of the ionosphere by the Appleton method would be exceedingly useful. The information acquired will be welcomed alike by astrophysicists interested in the solar chromosphere and by geophysicists whose chief concern is with the electrical condition of the earth's upper atmosphere.

Bolometric line contours. Most of the information we have concerning the constitution of stellar atmospheres is based ultimately on our knowledge of the processes involved in the formation of stellar absorption lines. The experimental foundation of this knowledge lies largely in the study of the contours or profiles of the lines. Research on this subject is occupying an increasing proportion of astrophysicists' time. Most of the work is photographic, i.e. the contours are derived from high-dispersion spectrograms by applying some sort of photometric measuring process to the plates after exposure and development. It is in this photometric examination that most of the troubles occur. Only those who have had to struggle with methods of photographic photometry realize how far from satisfactory is the photographic emulsion as a means for recording differences of light intensity. The difficulties are the more acute in astronomical work in that the general level of intensity to which the plates are exposed is so very much lower than in other branches of science. The experimental astrophysicist, then, will welcome any process which will avoid the use of the photographic emulsion for this work. Results of such a process are reported by Dahme[10], who has used a bolometric method for obtaining the total absorption of the infra-red hydrogen lines (Paschen series) in the solar spectrum. The number thus calculated of 3-quantum H atoms above each square centimetre of the photosphere agrees well with values obtained photographically. Similar measurements on the $4^2P \leftarrow 3^2D$ lines of Ca^+ near 8500 A. also yield values comparable with those obtained by independent methods. Various causes of distortion may, of course, influence the contours obtained by this as by any other method, but there seems little doubt that after the appropriate corrections have been applied,

the bolometric method can give as accurate results as the photographic, with considerably less trouble. There is probably little immediate likelihood of extending the process to short wave-lengths or to weak sources, but contour measurements in the infra-red solar spectrum alone are of sufficient importance to justify the method even if its use has to be confined to this limited field.

Automatic guidance of telescopes. For such purposes as stellar radial velocity measurements, where high-dispersion spectrograms with iron-arc comparisons are required, it is essential to use a slit spectrograph at the focus of the telescope rather than an objective prism. The image of the star produced by the objective must be accurately centred on the slit of the spectrograph, or the lines produced will be unsymmetrical. Inaccurate guiding can therefore not only produce spurious displacements of the spectrum lines by shifting their centres of gravity, but also render the spectrograms useless for contour measurements. Manual guiding during long exposures, however, requires constant and exacting attention from the observer, and towards the end of such exposures fatigue and resultant inaccuracy are almost inevitable. One method of getting over this trouble is exemplified by a device[11] recently brought into operation at Helsingfors. This apparatus records the accuracy of the guidance by photographing the star image on the slit whilst the spectrum is being recorded by the spectrograph itself. Though this device will enable unsuitable spectrograms to be rejected, it will not in itself, of course, improve the accuracy of setting. For this latter purpose an automatic photoelectric telescope guider[12] has been recently tried at Mt Wilson. The principal difficulty of design is the extremely small amount of light available to actuate the mechanism. Most photoelectric control devices work on 10^{-2} or 10^{-3} lumens, but an automatic telescope guider has to be satisfied with values ten million times smaller than this. The contrivance described can be operated by a star of magnitude 8·6 in a 60-in. telescope. A 90° roof-prism aluminized on its upper surfaces is set in the focal plane of the telescope and divides the light of a guiding star into two beams, which are fed alternately into a Zworykin electron multiplier. If the star image is not exactly centred on the knife-edge, the resulting flicker produces an alternating e.m.f. which is amplified and used to drive a correction motor until the flicker stops. The correcting impulse being proportional to the need for it, over-correction is eliminated. No mechanical relays are used, control being exercised entirely by electron tubes. A device of this kind will go far towards eliminating the routine work of an observatory, as well as towards improving the reliability of its spectroscopic results.

Liquid prism spectrograph. Both grating and prism spectrographs are in common use in astronomical observatories. The uses to which they are put are fairly sharply defined by their respective limitations: grating spectrographs are employed for bright objects to obtain high dispersion and resolving power, whilst prism spectrographs, being less wasteful of light, are used perforce for faint objects, though neither their dispersion nor their resolving power can in general compare with those of grating instruments. The division between these two functions is however being sapped from both sides. On the one hand, empirical knowledge of the effects

of different methods of ruling gratings, gained largely as a result of research in the extreme ultra-violet, is accumulating. Gratings can now be ruled in such a way as to throw most of the diffracted light into one particular order, thus making them more useful in low-intensity work. On the other hand, attempts to improve the dispersion of prism spectrographs date from very early times, with the introduction by Littrow of the auto-collimating type of prism mounting. Such mountings give good results when the effective aperture of the prism train is no more than a few centimetres, but with bigger prisms the want of homogeneity causes loss of definition. Large prisms of solid glass which will bear multiple transmissions are few and far between. An ingenious arrangement due to Evershed[13, 14], however, employs a hollow 63° prism of 9 cm. aperture filled with ethyl cinnamate, through which the light is transmitted 4, 6, 8, or even 10 times by using plane mirrors behind and in front. The liquid is not stirred, but is kept 10 ft. below ground for constancy of temperature. The dispersion with four transmissions at the *H* and *K* lines rivals that obtained in the third order of the 75-foot spectrograph at Mt Wilson, whilst the resolution is better than that in Rowland's solar spectrum map. The astonishing performance of this arrangement can best be gathered from the statement that measurements on a *print* with a foot-rule graduated in hundredths of an inch give wave-lengths correct to 0·006 A. in a 30 A. range near λ 4000, with only four transmissions. High-dispersion spectrograms of this kind will be invaluable in the study of line contours, of Zeeman effects, and of small shifts such as are caused by the Doppler, Einstein, and limb effects. The existence of liquids such as methyl naphthalene with still higher dispersive powers suggests that even better results may be hoped for.

Recording spectrohelioscopes. The work now being done internationally with the spectrohelioscope has been mentioned earlier (see *Bright solar eruptions and radio fading* above). This instrument has the great advantage over the spectroheliograph that it can follow rapid variations in the solar phenomena studied, but labours under the corresponding disadvantage that it produces no more accurate record of such variations than is provided by the observer's description of events. The history of astronomy, however, leaves no doubt of the importance of securing permanent records. Many of the most important discoveries can be attributed to the habit of preserving photographs after their immediate purpose has been served. It is encouraging, then, in view of the importance attached to continuous observation of the sun's surface in light of various wave-lengths, to learn that the problem of making the spectrohelioscope a photographically recording instrument is being tackled at several observatories. By attaching a motion picture camera to the visual instrument, Keenan[15] has studied the radial motions of prominences. The films obtained provide all the data necessary for the construction of the true three-dimensional trajectories of the knots of gas constituting eruptive prominences. The University of Michigan spectroheliokinematograph[16] which has been recently constructed is also used for prominence observations. Exposure times of individual pictures vary from 12 to 70 sec., and the intervening time-interval is 2·5 sec. By speeding-up the course of events many hundredfold in this way, the observer is

enabled to follow the formation, motion and destruction of prominences at leisure. At Mt Wilson, too, since May 1936, spectroheliograms have been taken continuously on 35-mm. motion picture film by an automatic recorder[8]. Here exposures lasting 50 sec. are separated by time-intervals of 4 min. The machine is started soon after sunrise and operates with very little attention until shut off an hour or so before sunset. Already, then, permanent records of the solar surface taken at short intervals in H and Ca^+ light are accumulating. None but the immediate results of this development can yet be foreseen, but its importance is obviously likely to be far-reaching. Future astrophysicists may have as good reason to be grateful for these records as present ones have for early stellar photographs and spectrograms.

Solar observations in the stratosphere. The spectroscopist has good reason for regretting that the earth's atmosphere at ground level absorbs completely all radiation of wave-lengths shorter than 1850 A.: the technique of vacuum spectroscopy in the laboratory, though vastly improved during the last decade, is likely to remain relatively laborious. How much of a handicap this absorption is to the astrophysicist can be realized from a single example. The strongest lines of the element most abundant cosmically, hydrogen, lie far out of reach in the extreme ultra-violet (the Lyman α line is at 1216 A.). But in addition to this absorption the astrophysicist has to put up with a further loss: of the ultra-violet radiation remaining, that below 2950 A. is intercepted by ozone molecules in the upper atmosphere. Thus the whole range of solar and stellar radiation below 2950 A. is beyond his reach if he observes from ground level. There remains, however, the possibility of taking photographs, at least of the solar spectrum, in the stratosphere. Several enquiries have recently been made into the practicability of experiments directed to this end.

In two of these[17] the solar radiation in space is regarded as equivalent to that of a black body at 6000° K.; then, using both micro-chemical data on the composition of the atmosphere at various heights attainable by sounding balloons, and spectroscopic values of the absorption coefficients of such gases as oxygen and ozone, it is possible to predict for each level the spectral energy distribution curve in the region 3000–2000 A. It is found that at heights above 15 miles there is reasonable hope for spectrographic observation in the range 2200–2000 A., i.e. between the Hartley band of ozone and the strong Schumann-Runge oxygen absorption. Such heights do not appear unattainable by manned balloons when recent successful ascents are remembered.

Another investigation[18], however, throws doubt on the validity of the assumptions on which this result depends. The apparatus used here for measuring the spectral quality of ultra-violet solar radiation in the stratosphere consists of a suitable combination of photoelectric cell and filter. The curves thus obtained suggest that the energy distribution outside the earth's atmosphere is by no means identical with that of a black body at 6000° K. The theoretical implications of this result alone are sufficiently important to justify attempts at spectroscopic verification.

In another discussion[19] going somewhat further than these, Saha considers

the possibility of using unmanned balloons for conveying spectrographs aloft. Sounding balloons capable of carrying a small spectrograph can reach a height of some 25 miles. There seems little doubt that this is well above the ozone layer, and therefore that a successful ascent would give a spectrogram showing the region 2950 to 1850 A. A spectrograph has actually been sent up in this way, but the surface which was used to diffuse light into the slit reflected poorly in this region. Another attempt, however, might well be successful in the near future.

It is also possible that at above 25 miles the stratosphere is sufficiently transparent to the region 1250 to 1000 A. for a measurable record of this radiation to be obtained on the plate with a reasonable exposure. This region is of special interest since it contains the resonance line of hydrogen, $H_L\alpha$. Theoretical calculations[20] of the width of this line—calculations based directly on the observed widths of the visible hydrogen lines—suggest that it should be of the order of a hundred times as wide as the red line $H_B\alpha$, which is already one of the broadest lines in the visible solar spectrum. The wings of such a tremendous absorption line would include the whole region 1400 to 1000 A. If a blank plate were obtained, then, we should be left in an extraordinary position of doubt. Would the failure mean that the balloon had not risen to heights penetrated by the radiation in question, or would it mean that such radiation never even reaches the earth, but is absorbed in the reversing layer of the sun? Co-operation between the astronomical spectroscopist and the geophysicist is required if these and many other allied questions are to be satisfactorily answered.

Astronomical applications of rotatory dispersion. Stellar spectra taken with objective prism spectrographs are useless for radial velocity work since comparison spectra cannot be photographed alongside them for the exact determination of wave-lengths. On the other hand, the spectra of very faint objects cannot be photographed with slit spectrographs since they are so wasteful of light. An interesting device[21] employing rotatory dispersion seems likely to make possible fairly precise radial velocity determinations from slitless spectrograms.

The apparatus was originally devised for the rapid photographic determination of colour indices. In front of the photographic plate at the telescope focus is put a combination of three different plane crystals. First, an Iceland spar crystal some 4 mm. thick gives by virtue of its double refraction two pencils of polarized light from each star. Behind this crystal is a quartz plate, cut perpendicular to its axis, which rotates the planes of polarization of these beams to a different degree for different colours. Finally, the light passes through a second Iceland spar crystal of the same thickness as the first, but turned through 90°. As a result of this arrangement, each star is recorded on the plate as four images. If the thickness of the quartz is suitably chosen and ordinary emulsions are used, two of these will be blue and two ultra-violet, their relative intensity depending entirely on the spectral type of the star. A *B* star will give strong ultra-violet images, an *M* star very weak ones. If one image of each colour is weakened a known amount by partial polarization (e.g. by using a tourmaline plate or even by the natural polarization in a Newtonian mirror), the photographic plate can be standardized by a method analogous to the grating method of determining emulsion characteristics. The great advantage of

this device is that all the images are produced in one exposure, which also serves to standardize the plate. The field is of course restricted by the size of the crystals used, but for colour index determinations in such objects as stellar clusters the method would seem to be of great promise.

The application to radial velocity work with slitless spectrographs can now be easily understood. A crystal system of the type described, used in conjunction with an objective prism, will give four spectra in the same way as with the objective alone it gives four direct images. Each spectrum will be crossed by dark bands (a great number will appear if a thick quartz plate is used), the intensity maxima in the two spectra from the "blue" images corresponding to minima in the spectra from the "ultra-violet" images. These bands can evidently be used as wave-length references. Perhaps the most accurate way would be to obtain microphotometer tracings of the two kinds of spectrum: their points of intersection will then serve as the wave-length standards. Clearly the method is not so precise as more direct ones in which a comparison spectrum is impressed on the plate beside the spectrum under examination. Nevertheless, for faint objects with great radial velocities, such as the extra-galactic nebulae, the arrangement would give fair accuracy whilst saving a great deal of exposure time. That something of the sort is badly needed will be realized when it is remembered that the faintest nebulae examined spectroscopically in the 100-in. telescope at Mt Wilson cannot be seen at the Cassegrain focus[22]. A direct photograph has first to be taken so that the observer can move the slit of the spectrograph predetermined distances in two co-ordinates from a neighbouring star which *is* visible.

References

(1) Merrill, P. W. *Publ. astr. Soc. Pacif.* **46**, 206 (1934); *Astrophys. J.* **83**, 126 (1936).
(2) Gleissberg, W. *Z. Astrophys.* **13**, 255 (1937).
(3) Beals, C. S. *Mon. Not. R. astr. Soc.* **96**, 661 (1936).
(4) Russell, H. N. *Mon. Not. R. astr. Soc.* **95**, 635 (1935).
(5) Swings, P. *Mon. Not. R. astr. Soc.* **97**, 212 (1937).
(6) Eddington, A. S. *Observatory*, **60**, 99 (1937).
(7) d'Azambuja, L. *C.R. Acad. Sci., Paris*, **204**, 1623 (1937).
(8) Richardson, R. S. *J. R. astr. Soc. Can.* **31**, 193 (1937).
(9) Newton, H. W. and Barton, H. J. *Mon. Not. R. astr. Soc.* **97**, 594 (1937).
(10) Dahme, A. *Z. Astrophys.* **11**, 93 (1935).
(11) Järnefelt, G. *Comment. phys.-math., Helsingf.*, **8**, 22 (1935).
(12) Whitford, A. E. and Kron, G. E. *Rev. sci. Instrum.* **8**, 78 (1937).
(13) Hargreaves, F. J. *Observatory*, **59**, 305 (1936).
(14) Evershed, J. *Observatory*, **60**, 29 (1937).
(15) Keenan, P. C. *Astrophys. J.* **83**, 55 (1936).
(16) McMath, C. and Pettit, E. *Observatory*, **60**, 85 (1937).
(17) Ramanathan, K. R. and Ramdas, L. A. *Proc. Ind. Acad. Sci.* **1** A, 308 (1934). Vassy, E. *Rev. Opt. (théor. instrum.)*, **15**, 81 (1936).
(18) Coblentz, W. W. and Stair, R. *Bur. Stand. J. Res., Wash.*, **17**, 1 (1936).
(19) Saha, M. N. *Bull. astr. Obs. Harv.* No. 905 (1937).
(20) Woolley, R. v. d. R. *Observatory*, **60**, 143 (1937).
(21) Öhman, Y. *Observatory*, **59**, 335 (1936); *Ark. Mat. Astr. Fys.* **25** B, No. 19 (1937).
(22) Hubble, E. *The Realm of the Nebulae.* O.U.P. (1936).

§ 3. VACUUM SPECTROSCOPY

By R. W. B. Pearse, A.R.C.S., Ph.D., Imperial College of Science and Technology

Introduction. Since the pioneer work of Schumann and Lyman in opening up the field of vacuum spectroscopy, progress has been rapid in pushing the limit of the explored ultra-violet spectrum to shorter and shorter wave-lengths. In recent years the introduction of improved glass gratings and the use of grazing incidence have enabled the limit to be pushed well into the range of wave-lengths where the methods of X-ray crystal spectroscopy are also applicable. With the improved methods of vacuum spectroscopy many spectra have already been investigated. In this report only a brief review of the progress made can be given; for a more detailed survey reference may be made to the recent book by H. Bomke[1].

Historical. Study of the ultra-violet spectrum has proceeded in steps. Early visual observations followed the solar spectrum to a point near the strong lines of ionized calcium in the region of 4000 A. In 1803 Ritter showed that sunlight possessed rays beyond this limit which were capable of blackening silver chloride. In 1842 Becquerel photographed the solar spectrum to about 3400 A. using a Daguerrotype plate. This is about the usual limit for glass instruments. Next, Stokes carried out investigations which made accessible the whole region of what we may call the normal ultra-violet spectrum. Using a fluorescent screen, quartz prisms and lenses and an electric spark as source, he appears to have reached the aluminium lines 1862, 1854 A. Subsequent improvements in quartz spectrographs and photographic plates have rendered investigation of spectra to about 2100 A. a comparatively simple matter.

Investigation of the region beyond this is complicated by three main factors. The one of these which gives this region its special distinction is the setting in of strong absorption by the oxygen of the air. The others are the difficulty of finding a suitable transparent substance to replace quartz, which becomes opaque, and the strong absorbing power of the gelatine of the photographic plate for these short wave-lengths. Schumann[2] overcame these difficulties by constructing a spectrograph which could be evacuated, by substituting fluorite for quartz and by making a special type of photographic plate[3] containing very little gelatine. He thus reached a limit later estimated as about 1200 A. Lyman[4] improved on this by substituting a concave grating for prism and lenses. By an ingenious arrangement of two slits he was able to measure wave-lengths in this region for the first time as well as to extend the known limit. In the case of helium he reached about 500 A. Millikan[5] and his collaborators using a high vacuum in the spectrograph, gratings specially ruled to throw as much light as possible into the first order for short wave-lengths and the "hot-spark" as source, attained 136 A. Meanwhile, the limit to which X-ray spectra from crystal gratings could be photographed was being pushed to longer wave-lengths, until, in 1925, Siegbahn and Thoraeus[6] obtained a record of lines at about 25 A. Attempts to fill in the intervening region were held up on the optical side on account of the very small reflecting power of the gratings for shorter

wave-lengths when illuminated at normal incidence, and on the X-ray side on account of the difficulty of finding crystals with sufficiently large grating space for use with longer wave-lengths. A new technique was necessary.

The first successful attempts to photograph this region were made by Dauvillier[7], using an extremely ingenious method. He deposited a thin layer of melissic acid a few hundred molecules thick on a flat lead plate. The molecules, which have a long chain structure, arrange themselves in a regular pattern which exhibits the properties of a grating. For melissic acid the grating space was found to be about 73·5 A. On account of experimental difficulties, Dauvillier was not able to make measurements of any great accuracy but he was successful in recording for the first time a few lines of oxygen, carbon, boron and thorium between 20 and 120 A. The method has now given place to less difficult and more accurate ones.

In 1923 A. H. Compton[8] showed that X rays were strongly reflected from plane surfaces of glass and silver, provided their incidence was almost grazing. Later, using a speculum grating in a similar manner, Compton and Doan[9] were able to obtain a diffraction pattern in three orders. In 1927 Osgood[10], using a concave grating at grazing incidence, was able to make a successful exploration of the region from 40 to 200 A. In the same year Thibaud[11], using a plane grating, obtained spectra from 50 to 170 A. Since these two researches there has been a steady improvement in spectroscopic technique in this region.

Spectrographs now employed for exploring the ultra-violet region from 2000 to 20 A. are, then, of three main types: prism instruments, instruments with gratings used at nearly normal incidence and instruments with gratings used at grazing incidence.

Prism vacuum spectrographs. With Schumann plates available and the absorption of the air overcome by evacuation of the apparatus, the main obstacle to the development of prism instruments is the difficulty of finding a satisfactory material for the optical parts. So far fluorite has been the most used. Good pieces of fluorite transmit very well to about 1600 A. and thereafter the transmission decreases steadily to a limit at about 1200 A. Quantitative measurements of the transmission of a piece of colourless fluorite from Zell in Baden have been made by Schneider[12] using a photographic method. A thickness of 1·9 mm. was used. Similar results for the same material were obtained by Powell[13] using a photoelectric method. The transmission is, however, very sensitive to small traces of impurities and large pieces exhibiting the best transmission are extremely rare. In view of this, great interest attaches to recent experiments on the growth of large single crystals of lithium fluoride in the laboratory, since lithium fluoride is superior to fluorite in being transparent to a limit of 1100 A. Schneider has carried out vacuum spectrographic measurements to determine the absorbing and reflecting powers and the refractive index of artificially produced lithium fluoride. In a short report[14] he states that the absorption coefficient increases slowly from 0·3 per cm. at 1600 A. to 1·0 per cm. near 1200 A. Below this it increases rapidly. Measurements of the refractive index of lithium fluoride had been made previously by Gyulai[15] as far as 2000 A. Schneider confirms these results and extends the observations to the

limit of transmission. At 1130 A. the refractive index has increased to 1·70. Further details are given in a longer paper[16]. The crystals were made by the method of Bridgman[17] by mechanically lowering a crucible containing the molten salt through the core of an electric furnace preferably evacuated. The time required for the formation and annealing of crystals about 3 cm. across was approximately 35 hr. Great purity of the material appears essential. The lithium fluoride used was prepared from carefully purified hydrofluoric acid and lithium nitrate. Schneider concludes that lithium fluoride in moderately thin pieces may be expected to transmit several per cent of the light down to wave-lengths as short as 1100 A. but this limit cannot be assumed for any piece chosen at random since impurities, poor polish and surface contamination may greatly reduce the transmission. Stockbarger[18] has also investigated the conditions under which large single crystals of lithium fluoride may be grown. He finds that great purity and a steep temperature gradient normal to the freezing layer are important. With a specially designed furnace, he was able to grow, regularly, single crystals 3 in. in diameter and of excellent optical quality, such that they could be cut and ground to any desired shape. The manufacture of these crystals is now being undertaken commercially.* It may also be noted in passing that even where grating spectrographs are used the lack of transparent solids is felt in the need for windows to keep harmful materials from the interior of the spectrograph and for filters to isolate various radiations for fluorescence and photochemical experiments.

A considerable number of fluorite spectrographs have been described. They differ mainly in the number of adjustments which may be made *in vacuo* and in the manner in which these are made. In the pioneer apparatus of Schumann provision was made for rotation of the slit about the collimator axis and adjustment of its width and length, for taking a comparison spectrum, for focusing the collimator and camera lenses, for adjusting to minimum deviation, for changing the inclination of the plate and for moving it in a vertical plane without disturbing the vacuum. With a 60° prism and two lenses of 16 mm. aperture and 15 cm. focal length Schumann obtained at 1400 A. a dispersion later estimated as 4 A./mm. A much simpler construction was described in 1919 by McLennan, Ainslie and Fuller[19] and this was followed in 1920 by a second construction by McLennan and Lewis[20]. As more recent examples an instrument used by Cario and Schmidt-Ott[21] for rapid work with small dispersion and one devised by Hese[22] for precise measurements may be cited. The latter instrument allows thirty spectra to be taken on a 9 × 12 cm. plate without breaking the vacuum; the dispersion is 8 A./mm. at 1854 A. with resolving power of some 30,200. Another instrument is described by Rose[22] in a companion paper. The Littrow type of mounting which has been so useful in the quartz region has been tried by Frerichs, whose instrument is described by Bomke[1].

Normal-incidence grating spectrographs. The concave grating may be used to investigate the region on the short-wave side of 2000 A. provided the air in the spectrograph is replaced by a less absorbing gas or is evacuated. For this purpose

* Harshaw Chemical Company, America.

it is desirable to use a type of mounting which is as compact as possible; therefore the mounting usually employed is that devised by Eagle where the slit is near the plate. This was the type of mounting used by Lyman[4] in the first successful investigation. In this instrument, Lyman employed a speculum metal grating ruled in 1903 at Johns Hopkins University with 15,000 lines to the inch and a radius of 97 cm. The diamond point was selected with the object of throwing as much of the light as possible into one order, in this case one first order. Two slits were provided so that one region of the spectrum could be directly compared with another of known wave-lengths; in this way wave-lengths in the new region could be measured. Later instruments have not departed much from this original design. Improvements have been mostly in the direction of improving the gratings themselves and in building instruments of greater range and dispersion. As examples of this two instruments may be mentioned. Compton and Boyce[23] have constructed a 2 m. normal-incidence vacuum spectrograph to cover, on one plate, the whole range from zero order to 2500 A. in the first order. The grating used in this is a glass one with 30,000 lines per inch and gives a dispersion of about 4·27 A./mm. Spectra have been obtained to wave-lengths as short as 300 A.; beyond this the intensity fell off, but with long exposures some lines were traced to a little below 200 A. The instrument was designed for the establishment of short wave-length standards and for the study of emission and absorption of gases. Harrison[24] has constructed a 21 ft. vacuum spectrograph, in which the grating of 6 in. aperture gave spectra over the range 2000 to 1000 A. with a dispersion of 2·7 A./mm. in exposures of about 1 hr. A high degree of resolution was obtained in the spectra of the elements of the first long period obtained from a hot spark as source. To save repeated exhaustion of the large volume, 830 l., of the body of the instrument a stopcock is inserted in the slit tube so that the source and slit may be changed as required and a gate serves to isolate the 40-l. camera box so that air may be let into this separately.

Grazing-incidence grating spectrographs. It has already been mentioned that in the region of very short wave-lengths the reflection from a grating is extremely small unless the light falls on it at very large angles of incidence. Spectrographs employing grazing incidence have come into very general use during the last few years. Both plane and concave gratings are used, and each has special advantages; either may be used for the study of the ultra-violet end of the optical spectra of atoms and molecules and for the study of very soft X rays. The concave grating was used by Osgood for pioneer work on soft X rays in the region of 40 to 200 A. and has been brought to a high degree of perfection by Siegbahn and his collaborators. A very clear description of this type of instrument is given by Ericson and Edlén[25]. Since the slit, grating and plate are very close together on the Rowland circle, the volume of a grazing-incidence spectrograph is much less than that of a normal-incidence spectrograph employing a grating of the same radius of curvature. The dispersion with a grating at grazing incidence is not uniform; at the extreme ultra-violet end it varies markedly with the angle of incidence. As the light becomes more nearly grazing the dispersion increases and becomes much larger than with

the grating at normal incidence. For example, a 1 m. grating with 30,000 lines/in. at an angle of incidence of 80° gives a dispersion of 1·58 A./mm. at 100 A. and 4·08 A./mm. at 1000 A., whereas at normal incidence the dispersion is about 8 A./mm. Thibaud used a plane grating for soft X rays at about the same time as the work of Osgood. Most early investigators used plane gratings for this purpose, as measurements were more accurate, adjustment was easier, and spectrum lines were sharp even with moderately wide slits. The plane grating at grazing incidence has also been used by Siegbahn and Magnusson[26] for the spectroscopy of ultra-soft X rays, M, N and O series. The same authors[27] describe a ruling engine constructed for ruling gratings for work in this region and show photographs of the results of microscopic examination of gratings for defects. They also describe the construction of a plane-grating spectrograph.

Atomic spectra. A series of recent investigations of atomic spectra in the extreme ultra-violet has been made by L. and E. Bloch. The spectrum of sulphur[28] has been extended to 328 A., and that of tellurium[28] to 157 A. Measurements of wave-length were made with reference to oxygen and are considered accurate to about 0·01 A. The lines were produced by means of an oscillating discharge. In a later paper the results of an investigation of the spectra of zinc and cadmium metallic vapours[29], excited by an oscillating electrodeless discharge, are given with assignments of lines to Zn II, Zn III, Zn IV, Cd II, Cd III and Cd IV. Both normal- and grazing-incidence vacuum spectrographs were used with a 1 m. grating. A third paper[30] extends Carroll's measurements of the mercury spectrum from 740 to 220 A. using a grating at grazing incidence. Carroll's measurements are also repeated and the complete table of wave-lengths given from 1727 to 222 A. An attempt was made to separate the lines according to degree of ionization. Boyce[31], with increased dispersion and resolving power, has revised and extended the extreme ultra-violet spectrum of xenon in the range 2000 to 600 A. Lists of classified lines are given for Xe I, Xe II, and Xe III. McLeod[32], exciting the spectrum of atomic iodine in a mixture of argon and iodine, has photographed 250 lines in the region to 1900 to 800 A. with a 1 m. grating of 30,000 lines to the inch. To prevent the passage of iodine into the spectrograph a continuous flow of pure argon was maintained from the spectrograph through the slit to the discharge tube. In the region of shorter wave-lengths 126 to 47 A. Edlén[33] has identified lines in the vacuum spark spectra of Ti, V, Cr, Mn, Fe and Co as due to combinations similar to those in Mg I; he has calculated the corresponding term systems. A doublet combination is also indicated which corresponds to Al I.

Standard wave-lengths. Weber and Watson[34] have recently determined wave-lengths of lines of C, N and O in the region 2317 to 1080 A. to serve as standards. The method used was that of overlapping orders, second-order lines of C and N being compared with first-order lines of Cu and Fe which have been measured with an interferometer. Very good agreement was obtained with previous measures by Bryce and Rieke, the mean deviation being 0·009 A. and the mean systematic deviation 0·001 A. The work was done with a 10-ft. normal-incidence vacuum spectrograph with a dispersion of 5·5 A./mm. Since the method of overlapping orders is

subject to small errors, More and Rieke[35] have redetermined the wave-lengths of several lines of C, N and O by another method. Using a 2 m. normal-incidence spectrograph, they have compared the first-order lines of these atoms with the first-order spectrum of Cu II. Certain lines of this spectrum, which have been computed with considerable accuracy by Shenstone from the combination principle, were used as standards. The values obtained were found to be in satisfactory agreement with those obtained by the other method and suitably weighted mean values are therefore put forward as wave-length standards for this region.

Molecular spectra. Beutler and Junger[36] have investigated the absorption spectrum of H_2 in the extreme ultra-violet region. At about 805 A. they find that the R and P branches become diffuse. The analysis shows that in the vibrational state $v=6$ of the $3p\pi\,^1\Pi_u$ terms, the rotational levels with $K'_d=1$ and 2 are sharp while those with $K'_d=3, 4, 5$ are diffuse. The rotational terms K'_e observed from the Q lines are all sharp. The phenomenon is explained as due to a process of auto-ionization $H^{\bullet}_2 \rightarrow H^+_2 + e$. For the energy of ionization a value 124429 $cm.^{-1}$ is obtained as compared with the value 124569 $cm.^{-1}$ obtained by Richardson and Davidson by extrapolation. Jeppesen[37] has photographed the emission spectrum of deuterium with a 2 m. grazing-incidence vacuum spectrograph. Constants for 37 bands of the $2p\,^1\Sigma \rightarrow 1s\,^1\Sigma$ system and for 29 bands of the $2p\,^1\Pi \rightarrow 1s\,^1\Sigma$ system are obtained. These are compared with the corresponding data for hydrogen. Electronic shifts of 4 and 23 $cm.^{-1}$ were observed for the systems respectively.

Of investigations of the spectra of heavier molecules mention must be made of a very interesting series of papers by Price. The spectra are observed in absorption with a grazing-incidence grating, an intermittent discharge from a condenser through a capillary being used as a source of continuum. In this way absorption is observed corresponding to transitions of the molecule from the normal ground state to several excited states, each different electronic transition, accompanied with various vibrational transitions, forming a different system of bands. In many cases these systems are found to form a Rydberg series leading up to ionization, and a very accurate value of the ionization potential may be obtained. Two types of electronic excitation may be distinguished. In the one the transition makes little change in the binding forces in the molecule; the rotational and vibrational constants are therefore little changed and the band system appears very symmetrical about the origin. In the other there is a marked change and the system appears strongly degraded. In the former case the non-bonding electron may be completely removed and a series of band systems obtained, converging to a limit. This limit will correspond to some state of the molecular ion. With acetylene Price[38] finds a rich absorption spectrum in the region 1520 to 1050 A. Two classes of bands are observed: in one the bands show P, Q and R branches and appear as narrow doublets with sharp heads degraded to the red; in the other they show only P and R branches and a single head. The fundamental band of each system was determined by comparing the spectra of C_2H_2 and C_2HD. Bands corresponding to transitions involving excited vibrational states show large vibrational isotope effects, whereas the band near the origin

shows only the small shift arising from the difference of the zero-point energies. The analysis gives two Rydberg series

$$\text{I: } \nu_0^n = 92000 - R/(n-0{\cdot}95)^2,$$

$$\text{II: } \nu_0^n = 92000 - R/(n-1{\cdot}5)^2, \quad n = 3, 4, 5, \ldots.$$

It is suggested that one of the valence electrons between the two carbon atoms is excited, the two series of terms arising from different couplings between the orbital angular momentum of the electron and the internuclear axis. Ethylene shows only one series. Price and Collins[39] have observed the spectrum of oxygen from 1300 to 650 A., using a 2 m. grazing-incidence vacuum spectrograph with a dispersion of about 2 A./mm. Ionization potentials of 16·1 and 18·2 V. are found with an error estimated to be not more than 0·1 V. Price[40] has also determined ionization potentials for methyl iodide, bromide and chloride by this method.

Duncan and Harrison[41] have studied the absorption of NH_3 in the region 1620 to 1450 A. with high dispersion of 1·3 A./mm. Duncan[42] has continued the work with mixtures of ND_3, ND_2H and NDH_2, using a smaller dispersion of 8·5 A./mm. for the region from 2200 to 800 A. Values of the frequencies for the isotopic molecules were calculated from the data for NH_3, taking into consideration changes of mass only and good agreement with the observed values was obtained. Duncan[43] has also worked with N_2O below 2200 A., finding some regions of discrete bands, some of continuous absorption. Ten electronic states were observed, some of which were fitted into a Rydberg series converging to an ionization potential of 12·66 V. Beyond 997 A. the absorption was continuous to about 850 A. Other researches on molecules include work on SO_2 and H_2S by Herreng[44] and by L. and E. Bloch and Herreng[45], and on CO by Tschulanowsky and Stepanow[46]. Rathenau[47] has studied the continuous absorption of light from a capillary spark discharge in the far ultra-violet region, 100 to 270 A. For the gases NH_3, N_2, N_2O, CO_2 and C_6H_6, he finds that the ratio of the absorption coefficients correspond approximately to those to be expected from addition of the individual atoms, without reference to their arrangement in the molecule. Schneider[48] has tried the effect of the addition of 80 per cent of N_2 on the absorption coefficient of O_2 in the maximum of the continuous band at 1450 A., using pressure less than 15 mm. of mercury; no alteration is found.

Absorption in solids. Schneider and O'Bryan[49] have been able to extend to the region beyond 1600 A. the measurements of Hilsch and Pohl on the absorption of alkali halides. Thin films of the salts were supported on plates of LiF 1 mm. thick, which, carefully chosen and properly cleaned, were transparent to 1050 A. The salt was distilled on to the plate within a Thibaud spectrograph. A number of bands were found with widths varying between 0·3 and 1·3 eV. Several other halides show only a series of sharp absorption edges with continuous absorption toward shorter wave-lengths.

Value of the charge on the electron. Among other results of vacuum spectroscopy mention should be made of the reconsideration of the traditional value of e. The closing of the gap between crystal grating spectroscopy and ruled grating spectro-

scopy has revealed a systematic discrepancy between the values of the wave-lengths obtained by the two methods, bigger than the errors of measurement to be expected. This could be removed if e were taken as $4{\cdot}805 \times 10^{-10}$ instead of $4{\cdot}774 \times 10^{-10}$. Since this has been discussed in a previous *Report* (**2**, 247, 1935) by Prof. H. R. Robinson no further details need be given here.

References

(1) Bomke, H. *Vakuumspektroskopie* (Leipzig, 1937).
(2) Schumann, V. *S.B. Akad. Wiss. Wien*, **2**A, 625 (1893).
(3) Schumann, V. *Ann. Phys., Lpz.*, **5**, 349 (1901).
(4) Lyman, T. *The Spectroscopy of the Extreme Ultra-Violet* (Longmans, 2nd ed. 1928).
(5) Millikan, R. A. and Bowen, I. S. *Phys. Rev.* **23**, 1 (1924).
(6) Siegbahn, H. and Thoraeus, R. *J. opt. Soc. Amer.* **13**, 255 (1926).
(7) Dauvillier, A. *J. Phys. Radium*, **8**, 1 (1927).
(8) Compton, A. H. *Phil. Mag.* **45**, 1121 (1923).
(9) Compton, A. H. and Doan, R. L. *Proc. nat. Acad. Sci., Wash.*, **11**, 598 (1925).
(10) Osgood, T. H. *Phys. Rev.* **30**, 567 (1927).
(11) Thibaud, J. *J. Phys. Radium*, **8**, 13 (1927).
(12) Schneider, E. G. *Phys. Rev.* **45**, 152 (1934).
(13) Powell, W. M. *Phys. Rev.* **45**, 154 (1934).
(14) Schneider, E. G. *Phys. Rev.* **47**, 789 (1935).
(15) Gyulai, Z. *Z. Phys.* **46**, 80 (1927).
(16) Schneider, E. G. *Phys. Rev.* **49**, 341 (1936).
(17) Bridgman, P. W. *Proc. Amer. Acad. Arts. Sci.* **60**, 305 (1925).
(18) Stockbarger, D. C. *Phys. Rev.* **49**, 200 (1936).
(19) McLennan, J. C., Ainslie, D. S. and Fuller, D. S. *Proc. roy. Soc.* A **95**, 316 (1919).
(20) McLennan, J. C. and Lewis, A. C. *Proc. roy. Soc.* A **98**, 109 (1920).
(21) Cario, G. and Schmidt-Ott, H. D. *Z. Phys.* **69**, 719 (1931).
(22) Hese, H., Rose, A. and Gräfin zu Dohna, *Z. Phys.* **81**, 745 (1933).
(23) Compton, K. T. and Boyce, J. C. *Rev. sci. Instrum.* **5**, 218 (1934).
(24) Harrison, G. R. *Rev. sci. Instrum.* **2**, 606 (1931).
(25) Ericson, A. and Edlén, B. *Z. Phys.* **59**, 656 (1930).
(26) Siegbahn, M. and Magnusson, T. *Z. Phys.* **88**, 559 (1934).
(27) Siegbahn, M. and Magnusson, T. *Z. Phys.* **62**, 435 (1930).
(28) Bloch, L. and Bloch, E. *J. Phys. Radium*, **6**, 441 (1935).
(29) Bloch, L. and Bloch, E. *Ann. Phys., Paris*, **5**, 325 (1936).
(30) Bloch, L. and Bloch, E. *Ann. Phys., Paris*, **6**, 561 (1936).
(31) Boyce, J. C. *Phys. Rev.* **49**, 730 (1936).
(32) McLeod, J. H. *Phys. Rev.* **49**, 804 (1936).
(33) Edlén, B. *Z. Phys.* **103**, 536 (1936).
(34) Weber, R. L. and Watson, W. W. *J. opt. Soc. Amer.* **26**, 307 (1936).
(35) More, K. R. and Rieke, C. A. *Phys. Rev.* **150**, 1054 (1936).
(36) Beutler, H. and Junger, H. O. *Z. Phys.* **100**, 80 (1936).
(37) Jeppesen, C. R. *Phys. Rev.* **49**, 797 (1936).
(38) Price, W. C. *Phys. Rev.* **45**, 843 (1934).
(39) Price, W. C. and Collins, G. *Phys. Rev.* **48**, 714 (1935).
(40) Price, W. C. *J. chem. Phys.* **4**, 539 (1936).
(41) Duncan, A. B. F. and Harrison, G. R. *Phys. Rev.* **49**, 211 (1936).
(42) Duncan, A. B. F. *Phys. Rev.* **50**, 700 (1936).
(43) Duncan, A. B. F. *J. chem. Phys.* **4**, 638 (1936).
(44) Herreng, P. *Rev. Opt. (théor. instrum.)*, **15**, 413 (1936).
(45) Bloch, L., Bloch, E. and Herreng, P. *C.R. Acad. Sci., Paris*, **203**, 782 (1936).
(46) Tschulanowsky, W. M. and Stepanow, B. I. *Phys. Z. Sowjet.* **10**, 292 (1936).
(47) Rathenau, G. *Physica*, **3**, 727 (1936).
(48) Schneider, E. G. *J. chem. Phys.* **5**, 106 (1937).
(49) Schneider, E. G. and O'Bryan, H. M. *Phys. Rev.* **51**, 293 (1937).

§4. COMPLEX ATOMIC SPECTRA

By E. W. Foster, B.Sc., A.R.C.S., Imperial College of Science and Technology

Mechanical aids to analysis. In the 1936 *Report* (3, 369) it was pointed out that the analysis of an atomic spectrum is carried out in two steps, the first of which consists in establishing a set of term values by the method of recurring wave-number intervals, and the second in the interpretation of the term scheme thus established. In the case of complex spectra, it may happen that much of the time spent on an analysis is taken up by tedious arithmetical work necessitated by the first step. A rough idea of the nature of this work is to be obtained from the following brief considerations.

The arithmetic involved in constructing a term scheme for a complex spectrum has often consisted essentially in systematically subtracting from each other the vacuum wave-numbers of component lines of a group for each of various groups of lines selected from the spectrum. Pairs of lines showing the same wave-number difference are collected; an allowable tolerance in the equality is discussed in the *Report* quoted above. Each pair, if it has the required significance, represents the combination of a term with two other terms; these are common to other significant pairs of lines exhibiting the same wave-number difference (interval). Many ways have been adopted in particular cases for obtaining significant pairs of lines suitable for providing a start on the analysis of a complex spectrum. It has often happened that a group of lines showing a tendency to reversal has been useful for this purpose; these may provide immediately a number of new terms derived from two terms represented by a repeated interval. When intervals between members of these new terms have been found to recur between other pairs of lines of a suitable group, it may be possible to establish further terms, and so on. The selection of groups of lines is guided by many considerations: regularities exhibited by the lines themselves, e.g. of spacing, of intensity, of contour or of structure, or knowledge of temperature class may be useful; in particular cases knowledge of isoelectronic, neighbouring or other spectra may be invoked. It has recently been emphasized[1] that, particularly in the case of very complex spectra, even when accurate wave-length measurements are available, the same difference may recur several times by chance[3] amongst various pairs of lines, and so fail to establish significant pairs and corresponding term separations.

Whatever detailed systematizations of this arithmetic may be adopted in the analysis of a complex spectrum, the process of determining and testing very many wave-number intervals is always required. A mechanical device described by Harrison is admirably suited for facilitating this type of work in many cases[2,3].

Harrison's arrangement is in two separate parts. One of these is a simple machine which is used for rapidly punching holes at measured distances along the central portion of a perforated paper tape, which is similar in shape to an ordinary cinematograph film; the punching process makes in it two more parallel rows of holes. One of these rows contains small rectangular holes having positions along the

tape corresponding, on some linear scale, to the wave-numbers of the spectrum lines. The other row consists of groups of circular holes corresponding, on the same scale, to certain regularly repeated groups of wave-number intervals; these holes form a system of calibration marks. Tapes up to 300 m. in length may be accommodated.

The second part of the arrangement, the interval sorter, is represented diagrammatically, with some mechanical simplification, in figure 1; corresponding simplifications are made in the description. The punched tape is stored on a reel A. It passes from A (figure 1 a) over a small roller into engagement with the lower half of a double sprocket wheel B, thereafter over a second roller and into a flat channel C, containing a central slot rather over 1 m. in length. A second similar roller and sprocket wheel arrangement at D may be used for isolating a loop E of tape, of constant length up to 20 m. The tape returns over itself along the channel to a second reel F, having engaged with the upper half of each sprocket wheel. A

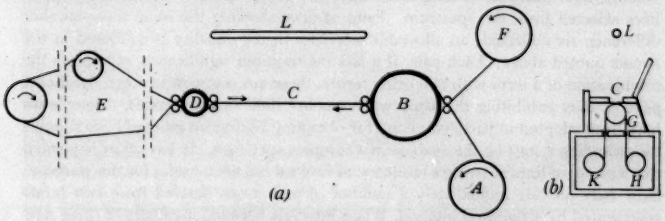


Figure 1. Interval sorter: (a) tape system, (b) recording system.

motor suitably geared to the sprocket wheels serves to drive the tape at constant speed from A to F. Photographic paper G (figure 1 b) 1 metre in width immediately underneath the channel C is driven with suitable constant speed at right angles to the direction of tape motion, from a roller H to a second roller K. A mercury arc lamp L is suspended over and parallel to the doubled tape.

When the machine, set up with punched tape and photographic paper, is in operation, a record will be made on the photographic paper whenever either two rectangular or two circular holes in the tape pass one another above this paper. Consider records formed with a fixed loop E of tape X metres in length, where this measurement is made from that edge of the photographic paper nearest the loop. The wave-number interval $\Delta\nu$ between two holes which produce a record at a distance x metres from this same edge is represented by $X+2x$ metres of tape. If this particular interval $\Delta\nu$ is repeated between other pairs of holes of the same type, then other records will be created in different parts of the paper, but they will all lie along a line distant x metres from the loop edge of the paper. When the paper is developed there will result a line of dots, approximately rectangular or circular in shape according to which set of holes has combined to produce the dots in question. Since tape and photographic paper move with fixed speed ratio, the position of any

dot on this line relative to an arbitrary fixed point on the line is linearly dependent on the wave-number ν corresponding to one member of the pair of holes responsible for the dot. The purpose of the circular holes is to produce (amongst a multitude of rectangular dots) an easily identifiable co-ordinate system of circular dots so that suitable ν and $\Delta\nu$ scales can be marked with ease, thus leading to the determination of any interval and the identification of the lines which produced it. Since any hole passes over all other holes of the same type lying between X and $X+2$ metres distant from it, all possible wave-number differences will be recorded whose magnitude lies in the region defined by X, $X+2$ metres of tape. Hence, by varying X on different runs all possible wave-number differences up to any desired magnitude are immediately made available. If rectangular holes of different sizes are punched to represent weak, medium and strong lines, it is possible to distinguish from the chart a dot caused by two medium lines from one caused by a strong and a weak line. It seems, therefore, that in many cases this arrangement will possess considerable advantages as regards speed, over other methods which have been used for aiding the determination of repeated intervals.

This machine provides equally well both real and accidental wave-number coincidences, and some interesting probability considerations concerning the elimination of the latter have been given by Harrison[3].

Albertson[4] has used this machine in connexion with an extensive analysis of the Os I spectrum. The instrument was used, first, for making a start on the analysis, when a selected group of about 180 wave-numbers likely to exhibit regularities were run through the machine and frequently recurring intervals ascertained; and, secondly, for reducing the time taken to comb a large wave-number list with a given interval. It was possible to determine from the record the approximate wave-number of a line involving a given interval, whereupon only a small number of check subtractions were necessary to determine the wave-number exactly, with a consequent large saving of time. The instrument has been used in connexion with analyses of Ir I, Sa I and Ce I (Ce I results not yet published), but there appears to be some slight doubt as to whether it will in general prove to be of very extensive value in connexion with rare earth spectra.

Atanasoff and Brandt[5] have suggested a device which may be added to a particular type of automatic computing machine so that the combination may be used for the same purposes as discussed in connexion with Harrison's apparatus. The machine in question is one of the punched-card tabulator type, such as are often used in statistical work. It is remarked that another type of punched-card tabulator which has recently become available may be used for sorting intervals without any additional apparatus. In most cases it appears that this type of apparatus is not likely to yield results with the rapidity of Harrison's, but with particularly complex spectra its characteristic of high accuracy may be an advantage. Atanasoff and Brandt's arrangement was developed on the Fe I spectrum, but no analyses in which it is stated that such instruments have been employed appear to have been published yet.

Automatic spectrogram comparator. The measurement and reduction to wave-

lengths of the many thousands of lines which may belong to a single complex spectrum is a formidable task. Harrison[6] has recently described an ingenious apparatus for automatically measuring and reducing spectrograms. Prof. Harrison has kindly informed the writer that one such apparatus now in use at Massachusetts Institute of Technology is recording wave-length measurements at the rate of 100,000 a week, and that when this instrument has been set up with a suitable calibration curve it is capable of measuring 1000 lines to an accuracy of 0·001 A. in 150 sec. The machine is complicated and no justice can be done to its mechanism in the short space available.

Rare-earth spectra. Albertson[1,7] summarizes some of the features which have been derived from the study of rare earth spectra. Only in the last few years has any significant progress been made in the analysis of these spectra, largely owing to their extreme complexity and to the inadequate differentiation between arc and spark spectra existing prior to the work of King, who began a detailed investigation of rare-earth spectra about 10 years ago; the observational material is not yet complete in all cases. The complexity of these spectra is illustrated by a remark made by Albertson in connexion with the analysis of the Gd I spectrum[7], that with strictest selection rules for L, S and J and parent terms, a fundamental electron transition in Gd I can yield more than 20,000 lines. If only the J rule is retained there is a possibility of 18 million lines, whereas the identical jump in Cs I produces two lines. Similar considerations will explain the feature, observed in most of the rare-earth spectra, that the number of lines obtained appears to depend only on the length of the exposure. King has compared the spectra produced in the arc, in the spark, and in the electric furnace at several different temperatures. It was found that the spectrum of the neutral atom was often very poorly developed in an arc source and relatively much more strongly in the furnace. Often the most prominent lines in the arc source belong to the singly ionized atom. The separation of lines of the neutral and ionized atoms was thus accomplished, and moreover extensive observations on the grouping of lines according to their response to different temperatures and excitation conditions in the furnace were made. Some other observations useful for grouping lines, and so aiding an analysis, were recorded at the same time; amongst these were characteristics of contour and hyperfine structure. King's lists usually contain detailed information at least on the most prominent lines of a spectrum.

Analyses of varying degrees of completeness have been made in the following cases (chronological order): Lu I, II, III[8], Yb I, II[9], Eu I[10], Eu II[11], Gd I[7], Sa I[7,12], Sa II[1]. It appears that the published analyses of Ce I, II are incorrect. The spectra of Lu are relatively simple compared with those of other rare earth spectra.

Albertson points out that the ground configurations of the neutral rare-earth atoms are now either known or can be predicted as follows ($6s^2$ is omitted from all configurations):

Atomic number	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Element	Ce	Pr	Nd	Il	Sa	Eu	Gd	Tb	Ds	Ho	Er	Tm	Yb	Lu
Configuration	$4f5d$	$4f^25d$	$4f^4$	–	$4f^6$	$4f^7$	$4f^75d$	$4f^85d$	$4f^{10}$	$4f^{11}$	$4f^{12}$	$4f^{13}$	$4f^{14}$	$4f^{14}5d$

The predictions were made by a detailed study of the characters of the various spectra as revealed by King's data. Most text-books discussing rare-earth spectra state that the normal configuration of the neutral atom is $4f^n5d6s^2$, where n increases progressively from 1 for Ce to 14 for Lu. It is now clear that there exists between $4f$ and $5d$ electrons the same type of variation of relative binding as exists between the d and s electrons in the long periods. The most intense lines of the spectra result from the transition $6s6p$ to $6s^2$, built on parent terms derived from the configurations given above.

Extremely interesting details of the hyperfine structure of Eu I have been discussed in the previous *Report* (**3**, 373).

References

(1) Albertson, W. E. *Astrophys. J.* **84**, 26 (1936). [Gives references to King's observational work.]
(2) Harrison, G. R. *Rev. sci. Instrum.* **3**, 753 (1932).
(3) Harrison, G. R. *Rev. sci. Instrum.* **4**, 581 (1933).
(4) Albertson, W. E. *Phys. Rev.* **45**, 304 (1934).
(5) Atanasoff, J. V. and Brandt, A. E. *J. opt. Soc. Amer.* **26**, 83 (1936).
(6) Harrison, G. R. *J. opt. Soc. Amer.* **25**, 169 (1935).
(7) Albertson, W. E. *Phys. Rev.* **47**, 370 (1935).
(8) Meggers, W. F. and Scribner, B. F. *Bur. Stand. J. Res., Wash.*, **5**, 73 (1930).
(9) Meggers, W. F. and Russell, H. N. Unpublished; see (1).
(10) Russell, H. N. and King, A. S. *Phys. Rev.* **48**, 1023 (1934).
(11) Albertson, W. E. *Phys. Rev.* **45**, 499 (1934).
(12) Albertson, W. E. *Phys. Rev.* **52**, 644 (1937).

§ 5. INFRA-RED SPECTRA OF ORGANIC MOLECULES

By Lotte Kellner, Dr.Phil., Imperial College of Science and Technology

Introduction. Though the results of the investigation of Raman spectra are included in this report in cases where their knowledge is necessary for complete analysis of the vibrational spectra, the title "infra-red" has been chosen, as the report is chiefly concerned with the research on infra-red spectra. Since the first *Report* (**2**, 342, 1935) appeared under this title, definite progress has been made both in the analysis of the spectra of complex molecules and in the obtaining of new experimental data on the infra-red spectra of organic molecules and their interpretation. For example, in the case of benzene and methyl alcohol the correct interpretation of the infra-red and Raman spectra has been made possible by the discovery of heavy hydrogen. The introduction of deuterium into organic molecules enables one to decide definitely, in which oscillations a group containing hydrogen takes part, and to what extent. The earlier paragraphs of the present report deal with the analysis of the vibration bands of certain molecules. The second part will be given over to a discussion of the new experimental data concerning the influence of the intermolecular and intramolecular forces on certain valence vibrations.

Analysis of the vibration spectra of C_6H_6 *and* C_6D_6. The almost complete analysis of the infra-red and Raman spectra of C_6H_6 and C_6D_6 has been accom-

plished by the team work of Angus, Bailey, Hale, Ingold, Leckie, Raisin, Thompson and Wilson[1,2]. Out of thirty-two bands appearing in the infra-red spectrum of liquid benzene (thirty-six in the Raman spectrum) it has been possible to assign twenty-five bands to definite modes of vibration. The unravelling of the benzene spectrum has been of great interest as it was hoped to decide the question of the molecular configuration of benzene. For the number of fundamentals and the coincidences between Raman lines and infra-red bands are connected with the symmetry of the molecule. The problem was, whether the benzene molecule possesses a centre of symmetry or not. Chemical evidence (equivalence of all six bonds with regard to substitution) tended to show that it had, while the Kekulé model (three alternating double bonds) gives a lower degree of symmetry to the molecule. It is known on theoretical grounds that in the case of a centrally symmetric molecule all vibrations appearing in the Raman spectrum should be inactive in the infra-red spectrum and vice versa. But there appear bands of the same frequency in both spectra and with surprisingly great intensity. The authors have shown, by a comparison of the absorption spectra of liquid and gaseous benzene, that the number of coinciding bands is smaller for the gas than for the liquid. This means that in the liquid state the benzene molecule is slightly distorted by the electric field of the surrounding molecules. This interaction seems to be sufficient to invalidate the selection rules, but not great enough to disturb seriously the modes of vibration. For, if this were the case, the frequencies of the vibrations of the gaseous and liquid benzene could not coincide within a few wave-number units, as they do.

Bright Wilson[3] had shown, with the help of group theory, that benzene (plane regular hexagon) should have four infra-red active fundamentals, three of them twofold degenerate, and seven Raman-active fundamentals, five of them degenerate. Of the remaining eleven, which are inactive in both spectra, two are degenerate. Wilson has furthermore given the thirty modes of vibration and the formulae for the connexion between the force constants and the frequencies. He used a potential function of the usual kind, in which the potential energy of the molecule is a quadratic function of the displacements of the atoms from their equilibrium positions, and assumed that equilibrium is maintained by three forces acting on the carbon atoms, and three forces acting on the hydrogen atoms: namely, a valence force between the valence bonds, a force which maintains the angles between the linkages, and a force which prevents the twisting of the bond out of the plane of the molecule. No interaction has been taken into account.

In order to assign the observed bands to the different fundamentals, Bailey, Hale, Ingold and Thompson[1] applied Teller's product theorem[4] which connects the fundamentals of C_6H_6 with those of C_6D_6 of the same symmetry class, on the assumption that in C_6H_6 and C_6D_6 the forces and interatomic distances are exactly the same. The term "symmetry class" refers to the type of symmetry of the modes of vibration. It was assumed that benzene had the symmetry of a plane regular hexagon. With the help of this theorem the following fundamentals (table 1) were assigned to these molecules:

Table 1. Fundamental frequencies ($cm.^{-1}$) of C_6H_6 and C_6D_6.
(Notation by Bright Wilson)

	Infra-red active vibrations				Raman-active vibrations					
	Non-degenerate	Degenerate			Non-degenerate	Degenerate				
	ν_{11}	ν_{18}	ν_{19}	ν_{20}	ν_1 ν_2	ν_6	ν_7	ν_8	ν_9	ν_{10}
C_6H_6	671	1030	1485	3080	992 3062	606	3047	1596	1178	849
C_6D_6	503	813	1333	2294	945 2292	577	2264	1559	867	661

Of these vibrations ν_1 is the so-called breathing frequency of the benzene molecule, in which all carbon atoms move symmetrically inward and outward. In ν_{10} and ν_{11} mainly a bending of the C—H bonds out of the plane of the ring takes place. In ν_2 and ν_{20} the C—H bonds are stretched. It will be seen from the table that some frequencies, namely those in which chiefly the C—H bonds are vibrating, are changed very much when the hydrogen atoms are replaced by deuterium, while some are changed only to a small extent. From the successful analysis of the vibration spectra, it can be concluded that the benzene molecule has the configuration of a plane regular hexagon. Kohlrausch[5] has evaluated the force constants from Wilson's formulae. His results are given in table 2 together with those of Lord and Andrews[6] who used the same formulae but corrected a small error of Kohlrausch.

Table 2. Force constants for C_6H_6 and C_6D_6 (values in 10^5 dyne/cm.)

Force	Constant	C_6H_6		C_6D_6	
		Kohlrausch	Lord and Andrews	Kohlrausch	Lord and Andrews
C—C stretching force	K	7·65	7·58	7·58	7·62
C—H or C—D stretching force	q	5·02	5·05	5·10	5·11
C—C—C bending force	k	0·64	0·65	0·64	0·65
C—C bond torsion	κ	—	0·23	—	0·23
C—H or C—D bending out of plane of three C atoms	h	0·34	0·34	0·34	0·34
C—H or C—D bending in plane of molecule	H	0·74	0·76	0·86	0·78

It should be noted that the C—C stretching force constant is very nearly the average (7·38) between the constants for C—C single bonds (4·96) and C=C double bonds (9·79)[7]. This confirms the idea of the theoretician that in the benzene molecule twelve of the available eighteen C-electrons are used for the six single bonds between the carbon atoms while the remaining six are shared uniformly by the carbon linkages. It is possible to compute the inactive vibrations from Wilson's formulae. In table 3 a comparison is given between Bailey's, etc.[1], Kohlrausch's[5], Lord and Andrews'[6] computations and those frequencies of benzene which appear in its electronic spectra[2] and as forbidden lines in the infra-red and Raman

spectrum. In the last column are given some bands of pyridine C_5H_5N whose spectrum shows a surprising agreement with that of benzene, though it no longer possesses a centre of symmetry.

Table 3. Frequencies ($cm.^{-1}$) of C_6H_6, C_6D_6 and C_5H_5N

Frequency number	Computed						Observed					
	Kohlrausch		Lord and Andrews		Bailey and Ingold		Electronic spectra		Raman and infra-red		C_5H_5N	
	C_6H_6	C_6D_6	C_6H_6	C_6D_6	C_6H_6	C_6D_6	C_6H_6	C_6D_6	C_6H_6	C_6D_6	Raman	Infra-red
3	1258	1058	1190	930	1240	940	1219	978	1291	978	1207	1210
4	—	—	538	395	—	—	—	—	—	—	—	—
5	—	—	1520	1480	—	—	—	—	—	—	—	—
12	1000	948	1008	960	766	730	—	—	—	—	—	—
13	3056	2296	3063	2294	3062	2294	—	—	—	—	—	—
14	1852	1842	1854	1844	1720	1640	—	—	—	—	—	1924?
15	1094	838	1145	816	1145	820	—	—	—	—	1144	1140
16	?	?	406	314	—	—	395	—	—	—	401	—
17	?	?	1160	1070	890	780	—	—	—	—	—	—

Lord and Andrews arrived at their results by considering the repulsion between two similar neighbouring atoms (C or H), which varies from one vibration to another and which causes a corresponding change in the force constant H. Against the comparison with pyridine the objection may be made that the modes of vibration of this molecule will be different from those of the C_6H_6 or the C_6D_6 molecule.

Analysis of the vibration spectrum of methyl alcohol. In the case of CH_3OH Bartholomé and Sachsse[8] have studied the frequency shift which occurs when heavy hydrogen is substituted for ordinary hydrogen in the OH group. They found that the band 3400 of CH_3OH experiences the greatest isotopic shift of all observed bands to 2420 for CH_3OD, which makes it quite certain that this band is associated with the valence vibration of the OH linkage. Very interesting in this connexion is the fact that, while the intensity of the OH band is 100 (in arbitrary units), the OD vibration has only an intensity of 20. It has been possible to determine the other fundamentals of CH_3OH: the band 1025 corresponds to the oscillation of the OH group against the CH_3 group; 3000 and 2943 are the antisymmetric and the symmetric valence vibrations of the CH bonds; 1470 and 1360 are caused by the deformation vibrations of the CH_3 group perpendicular and parallel to the symmetry axis of the group. The remaining frequencies 695 and 725 are assigned to modes of vibration, in which the CH_3 group executes torsional oscillations with regard to the OH group, the first one in the HOC plane, the second one perpendicular to this plane. From the equal intensity of these two bands follows that the angle HOC is different from 180°. No free rotation about the C—O bond seems to occur.

Near infra-red spectra of organic molecules and intensities of overtones. No attempt has been made to extend the interpretation of the spectrum of methyl alcohol to the overtones and combination tones of the fundamental vibrations,

while in the case of benzene it has been possible to assign eleven bands to binary combinations. Difficulties arise in the interpretation of this part of the spectrum, because the selection rules for the overtones of complex molecules are not known, except for a few molecules of relatively high symmetry. But the great similarity in the spectra of all organic molecules between 0·8 and 3μ. makes it certain that the same groups cause the bands in this region, i.e. the CH_3, CH_2, C=O, O—H and N—H groups are principally responsible for the near infra-red bands. For some time it has been assumed that the 1·7μ. band of organic molecules was the first overtone of the CH vibration at 3·4μ. but Kellner[9] has shown that, in the case of the aldehydes at least, it seems more probable that this band is a combination of the C—H valence vibration and the parallel deformation frequency of the CH_3 groups. The 1·7μ. band of the aldehydes is accompanied by the first overtone of the C—H valence vibration on the long wave-length side and by the third overtone of the perpendicular deformation vibration on the short wave-length side. The combination frequency is the most intense of the three bands; in the case of a centrally symmetrical molecule it would be the only frequency of non-vanishing intensity. For the aldehydes it has been possible to show that the bands between 1 and 2·1μ. can be interpreted as overtones and combination tones of four fundamentals only.

As the intensity of the bands is proportional to the number of oscillators, the absorption would be expected to increase with the number of similar vibrating groups. Unfortunately, the absorption of the fundamental bands is usually so strong that it is not possible to make quantitative measurements of the absorption of the fundamentals; the quantitative absorption measurements have been confined to overtones and combination tones. The total absorption of a band, i.e. the integral of the absorption coefficient over the frequency breadth of the band, should increase proportionally to the number of oscillators. Timm and Mecke[10] have observed the total absorption of the third overtone of the C—H vibration of some halogen derivatives of ethane, methane and ethylene, and found that the ethane derivatives followed the expected law much better than the methane derivatives. For the paraffins, olefines and acetylenes on the other hand, Gänswein and Mecke[11] showed that the intensity of absorption of the third overtone increases proportionally with the number of C—H valences of equal neighbours. It is necessary to study an overtone which is free from overlapping by another overtone or combination tone, as, otherwise, resonance between the two vibrations is likely to occur.

Influence of intramolecular forces on valence vibrations. The intramolecular interaction has been studied by examining the frequency shift of one particular band in a homologous series and by varying the neighbouring groups of the oscillator under consideration (an oscillator in this sense is an OH, CH or NH bond, etc.).

Investigations of the behaviour of the NH and OH vibrations in various derivatives have been made by Wulf and Liddell[12], Wulf, Liddell and Hendricks[13], M. and R. Freyman and Rumpf[14,15], Barchewitz[16], Naherniac[17] and others. Cheng and Lecomte[18] have demonstrated the influence of various neighbouring groups on the C—Cl vibrations, and Gillette[19,20] has studied the C=O band, when the components of the molecule are varied.

Wulf and Liddell and Hendricks investigated their specimens in a very dilute solution (0·01 molal), so that the influence of neighbouring molecules is negligible. As they used a very high resolution (6 A. effective slit-width), they were able to resolve the bands into their components. No indication of any rotational structure was observed; the half-breadth of the first overtone of the characteristic OH and NH vibrations was only of the order 35 to 40 cm^{-1}. They found that in *ortho*-derivatives, where the substituent is adjacent to the OH or NH group, the OH or NH band was shifted to greater frequencies as compared with the unsubstituted compound if the substituent was a methyl group, but to smaller frequencies if the substituent was a halogen or an NO_2 group. In *meta*-derivatives, where the distance between the groups is increased, the shifts are in the same directions as in the *ortho*-derivatives but smaller in magnitude. In *para*-derivatives (two groups at opposite positions in the benzene ring) the substitutions produce no measurable shifts. For phenol C_6H_5OH the authors found only one frequency at 7050, while for *o*-chlorophenol C_6H_4ClOH they got two, 7050 and a much stronger component at 6910. For catechol ($C_6H_4(OH)_2$, with two OH groups in adjacent positions) two bands of nearly equal intensity were observed at 6970 and 7060, while resorcinol (*m*-$C_6H_4(OH)_2$) showed only one band at 7050 of double the intensity. Pyrrole C_4H_4NH (five-membered ring with two alternating double bonds) gave two bands at 6800 and 6840, the latter much weaker than the former.

As regards the influence of *ortho*-, *meta*- and *para*-substitution on the original position of the NH vibration, these results have been confirmed for the second overtone of the NH vibration near 10,000 cm^{-1} by M. and R. Freyman and Rumpf. But Barchewitz has shown that phenol (either pure or dissolved in CCl_4) possesses two bands at λ7466 A. and λ7598 A. corresponding to the third overtone of the OH frequency. In *o*-chlorophenol the bands are both shifted to smaller frequencies as compared with phenol and the ratio of intensities is reversed. In *p*-chlorophenol the bands are shifted again to smaller frequencies but by smaller amounts and the ratio of the intensities is the same as in phenol.

Naherniac used a grating spectrometer of high resolution (3 A. effective slit-width) between 0·8 and 1·2μ. and studied the OH band between 0·96 and 0·98μ. for alcohols, phenols and fatty acids in the liquid and gaseous state. The results of his comparison of the frequencies of the liquid and gaseous state will be given in the next section. For phenol (either pure or dissolved in CCl_4) he found only one band, which is contrary to Barchewitz's results and in agreement with Wulf and Liddell's; but on the other hand he could find no doubling of the *o*-chlorophenol band. The fatty acids showed an OH vibration in the gaseous state. For the first three alcohols of the homologous series $C_nH_{2n+1}OH$ the OH band decreased steadily in frequency, and remained at the same position from $n=3$ onwards.

Cheng and Lecomte investigated the behaviour of the C—Cl vibrations in different substances between 640 and 1400 cm^{-1}. They observed that monohalogenated compounds possessed one frequency characteristic of the C—Cl valence, while molecules containing the group CCl_2 showed three frequencies corresponding to the three fundamentals of a triatomic molecule. The characteristic C—Cl

valence vibration shifted steadily to greater frequencies when the adjacent groups were changed from CN to COH, COCl, COOH and COOR, while the CCl_2 deformation frequency decreased in the same direction. Substances Cl_3—C—*R* containing the group CCl_3 showed no set of frequencies which could be identified with the frequencies of a pyramidal model XY_3. It seems that in this latter case the influence of the radical *R* is so great that the interaction between CCl_3 and *R* becomes of the same order as the forces causing the inner vibrations of the CCl_3 group, while in the two former cases the interaction is only of the second order as compared with the forces which hold the C—Cl and C—Cl_2 groups together. Cheng and Lecomte succeeded in identifying, besides, one C—C valence vibration, which shifted to higher frequencies the more halogen atoms were introduced into the molecule. For molecules with a given number of halogen atoms the C—C frequency decreased in the same way as for the C—Cl valence vibration, when the radical *R* was varied.

Gillette's work was concerned with the C=O vibration at approximately 1720 cm^{-1}. He found the C=O vibration in acetic acid CH_3COOH at 1721. The introduction of a halogen into the molecule shifted it to higher frequencies; the shift amounts to about 15 cm^{-1} per Cl atom and 9 cm^{-1} per Br atom. The introduction of iodine has no measurable influence on the position of the band. The substitution of methyl or ethyl groups for the hydrogen atoms of the CH_3 group in the acetic acid has the reverse effect; the frequencies become smaller.

Pauling[21] has given an explanation of the results of Liddell and Wulf on the splitting of the OH and NH bands in *o*-chlorophenol, pyrrole, etc. He assumed an interaction of the electric charges on the end atoms of the vibrating bonds. In the instance of *o*-chlorophenol, the chlorine atom appears negatively charged, while the hydrogen of the OH bond carries an excess of positive charge. Consequently, an attraction occurs between the Cl and the H, and the result is a quantum-mechanical resonance between two molecular forms, a *cis*-form, in which the OH bond is turned toward the C—Cl bond, and a *trans*-form, in which it is turned away from that bond. The *cis*-position corresponds to a weakening of the OH valence bonds, i.e. to the lower frequency 6910 (lower potential energy, stable form), the *trans*-position to a strengthening of the OH bond and the higher frequency 7050. The intensity ratio 10 : 1 of the *cis*- to the *trans*-vibration indicates that 91 per cent *o*-chlorophenol are in the *cis*-form and 9 per cent in the *trans*-form. For catechol (two adjacent OH groups) a *cis-trans*-configuration is assumed; i.e. one H atom is turned away from the neighbouring OH linkage (repulsion between two positively charged H atoms), and the second H atom is turned towards the oxygen of the first OH bond (attraction between negatively charged hydrogen and positive charge of oxygen). To the *cis*-OH bond corresponds the OH frequency 6970, to the *trans*-configuration the frequency 7060. From the existence of two bands for pyrrole Pauling concluded that this molecule exists in two forms: a coplanar form, in which the hydrogen of the NH bond lies in the plane of the ring, and a non-coplanar form. For pyrrole M. and R. Freyman confirmed this interpretation, but for phenol and *o*-chlorophenol discrepancies occurred with the

measurements of Barchewitz and Naherniac, as mentioned above, so that it seems doubtful whether the doubling of the *o*-chlorophenol band may not be due rather to a resonance with another overtone than to the existence of two molecular configurations. A study of the fundamental of the OH vibration of *o*-chlorophenol would solve this problem.

Naherniac, on the other hand, has given an interpretation which is equally applicable to all the phenomena mentioned above. He assumed intramolecular van der Waals forces which are responsible for the interactions between the linkages. Three effects have to be taken into account: the induction of electric moments by the permanent dipole moments of the bonds, the polarization effect and the dispersion effect. The induction and dispersion potential are always negative; i.e. they cause a weakening of the valence force, while the polarization potential can have a positive or negative sign. It is obvious that, as these intramolecular forces depend on the interatomic distances, the interaction will be the largest in *o*-derivatives and almost negligible in *p*-derivatives. If the attractive forces prevail, the result will be a weakening of the valence bond with a consequent shift to smaller frequencies.

All these intramolecular interactions, as considered above, may be termed second-order effects, for it follows from the change in the accompanying frequencies that the corresponding valence forces are changed only by about 1 to 13 per cent. But there are cases known in which the intramolecular interaction becomes so great (of the order of the valence forces), that a new bond is formed and the former valence vibration disappears[22, 23, 24, 25]. These phenomena are known to the chemist under the name "chelation", which means that a cyclization takes place between atoms of adjacent linkages, in which always one of the binding atoms is hydrogen, and it is bound to either oxygen or nitrogen. The nature of this linkage is still rather obscure.

Influence of intermolecular forces on valence vibrations. The intermolecular forces depend on the distances between neighbouring molecules. These distances can be altered by raising the temperature of the substance or by dissolving it in another substance.

The influence of these forces on the valence vibrations has been studied by investigating the infra-red spectra of liquids at different temperatures or of solutions of varying concentration. These forces are of the same kind as the intramolecular forces, and they are caused by the same interaction effects[14]. We have here the same transitions from interaction of the first order (formation of "cyclized" molecular aggregates with corresponding disappearance of the valence vibration of the bond participating in the cyclization) to second-order interaction, where only a shift of the frequency occurs.

Naherniac[17] has shown that an interaction of the first order takes place in the COOH group of the fatty acids: here we have an OH band of the substance in the gaseous state which is replaced by continuous absorption for the liquid state. The C=O frequency of acetic acid is shifted from 1740 to 1786 when the temperature is raised from 25 to 172° C.[22]. At intermediate temperatures both bands are present in the spectrum, indicating that both molecular forms (simple molecules and asso-

ciated molecules) occur. In the case of the alcohols and phenols[17, 26, 27, 28] both OH vibrations appear in the spectrum of the liquid: the higher frequency of the free molecule and the lower frequency of the molecule associated with the surrounding molecules by the intermolecular forces. The OH band of the free molecules is very weak in a pure liquid (small number of free molecules), but grows more and more intense, when the temperature is raised or when the liquid is diluted with a neutral liquid as CCl_4. Here the difference in the frequencies corresponds to a difference between the OH valence forces of about 15 per cent (second-order effect). In all these cases the CH bands were quite unaffected by the rise in the temperature or by the dilution with another liquid. This fact makes it quite certain that the associated molecules are linked together through the COOH and OH groups.

References

(1) Angus, W. R., Bailey, C. R., Hale, J. B., Ingold, C. K., Leckie, A. H., Raisin, C. G., Thompson, G. W. and Wilson, C. L. *J. chem. Soc.* pp. 925, 931, 966, 971 (1936).
(2) Ingold, C. K. and Wilson, C. L. *J. chem. Soc.* pp. 941, 955 (1936).
(3) Bright Wilson, E. *Phys. Rev.* **45**, 706 (1934).
(4) Teller, E. Unpublished, see last paper in (1).
(5) Kohlrausch, K. W. F. *Phys. Z.* **37**, 58 (1936).
(6) Lord, R. C. and Andrews, D. H. *J. phys. Chem.* **41**, 149 (1937).
(7) Sutherland, G. B. B. M. and Dennison, D. M. *Proc. roy. Soc.* A, **148**, 250 (1935).
(8) Bartholomé, E. and Sachsse, L. *Z. phys. Chem.* B, **30**, 40 (1935).
(9) Kellner, L. *Proc. roy. Soc.* A, **157**, 100 (1936).
(10) Timm, E. and Mecke, R. *Z. Phys.* **81**, 363 (1935).
(11) Gänswein, G. and Mecke, R. *Z. Phys.* **99**, 189 (1936).
(12) Wulf, O. R. and Liddell, V. *J. Amer. chem. Soc.* **57**, 1464 (1935).
(13) Wulf, O. R., Liddell, V. and Hendricks, S. B. *J. Amer. chem. Soc.* **58**, 2287 (1936).
(14) Freyman, M., Freyman, R. and Rumpf, P. *C.R. Acad. Sci., Paris*, **203**, 309 (1936).
(15) Freyman, M. and Freyman, R. *J. Phys. Radium* **7**, 476 (1936).
(16) Barchewitz, P. *C.R. Acad. Sci., Paris*, **204**, 1184 (1037).
(17) Naherniac, A. *Ann. Phys., Paris*, **7**, 528 (1937).
(18) Cheng, H. Ch. and Lecomte, J. *Ann. Phys., Paris*, **5**, 427 (1936).
(19) Gillette, R. H. and Daniels, F. *J. Amer. chem. Soc.* **58**, 1139 (1936).
(20) Gillette, R. H. *J. Amer. chem. Soc.* **58**, 1143 (1936).
(21) Pauling, L. *J. Amer. chem. Soc.* **58**, 94 (1936).
(22) Errera, J. and Mollet, P. *J. Phys. Radium* **6**, 281 (1935).
(23) Freyman, M. and Freyman, R. *J. Phys. Radium* **7**, 506 (1936).
(24) Freyman, M. and Freyman, R. *C.R. Acad. Sci., Paris*, **202**, 312 (1936).
(25) Freyman, M. and Freyman, R. *C.R. Acad. Sci., Paris*, **202**, 1850 (1936).
(26) Errera, J. and Mollet, P. *Nature, Lond.*, **138**, 882 (1936).
(27) Freyman, R. *C.R. Acad. Sci., Paris*, **204**, 1063 (1937).
(28) Fox, J. J. and Martin, A. E. *Nature, Lond.*, **139**, 507 (1937).

X-RAY ANALYSIS AND APPLICATION OF FOURIER SERIES METHODS TO MOLECULAR STRUCTURES

By J. MONTEATH ROBERTSON, M.A., D.Sc., Ph.D.
Davy Faraday Laboratory, Royal Institution, London, W. 1

§ 1. INTRODUCTION

The use of X rays as a tool for investigating the structure of matter, which began with Laue's discovery of the diffraction of X rays by crystals in 1912, has now grown into a subject of considerable dimensions, and even in the last few years several thousands of original papers have been published. This great development of the subject is partly due to the variety of materials, ranging from metals to complex biological structures, which display a sufficient degree of order in their molecular arrangement to give the diffraction effect.

Several aspects of X-ray analysis have already been mentioned in these *Reports*. The present article aims at giving some account of the X-ray method at its full power, i.e. when both the intensity and the position of the diffracted beams are utilized in building up a picture of the arrangement and distribution of the ultimate scattering units. In general (although there are important exceptions) this can only be done when single crystals of the material are available. Although most solid matter is found to display certain crystalline properties, the structure does not always present a completely regular array of molecules. The regularity may be confined to one or two dimensions, or to small particles which have a random orientation with respect to each other, and these factors tend to limit the amount of information which can be obtained by X-ray analysis. In these cases of incompletely crystalline bodies we can usually regard certain sets of X-ray reflections as being suppressed or superimposed on others. It is thus of importance to understand the analysis of single, complete crystals before studying the less perfect materials.

In this article it is not possible nor is it desirable to present all the developments of the subject of X-ray analysis in historical sequence. We give instead a brief account of the development of the subject in §2, in which reference will also be found to certain aspects which it is not possible to treat in detail, such as the geometrical side of crystallography and space-group theory, and the interpretation of X-ray crystal diagrams. This is followed in §3 by an account of the physical theory and formulae now used in quantitative work. In the later sections certain new developments are mentioned and descriptions are given of some important structure determinations recently carried out. These have been chosen to illustrate the methods described in §3, and also because of their bearing on physical and chemical problems of molecular structure. No attempt has been made to summarize or even mention all the structure determinations of recent years. For this the reader

is referred to the *Strukturbericht* vols. 1–3 by Ewald, Hermann and others[97] and to the *Annual Reports of the Chemical Society*. It may be added that although X-ray studies of crystal structure are very numerous, the exact solution of a complicated structure by making the fullest possible use of the X-ray data is a lengthy operation, and has only been carried to completion in comparatively few cases.

§ 2. DEVELOPMENT OF X-RAY ANALYSIS

Mathematical crystallography. Before the discovery of the diffraction of X rays by crystals, and indeed before the discovery of X rays at all, the mathematical theory of crystallography was fairly complete. The early work of Hauy, Frankenheim, Bravais, Sohnke and others was extended and completed between 1885 and 1894 by the work of three independent investigators, Fedorov in Russia, Schoenflies in Germany and Barlow in England. Their great achievement lay in working out all the geometrically distinct ways of arranging the symmetry elements which are possible in a crystal. These symmetry elements are limited by the law of rational indices. Thus it can be shown that in addition to certain operations involving translations and reflections, only two, three, four, and sixfold axes are possible; and all the ways of combining such elements can be studied, most conveniently by the theory of groups. The mathematically distinct arrangements which are possible, known as the "space groups", are found to number 230, and every crystal must conform to one of these groups, although it does not follow that every group is physically possible. An excellent treatment in English of the whole subject has been given by Hilton[51].

The most convenient way of regarding any crystal is to think of it as consisting of one or other of these possible space arrangements of symmetry elements. Any molecules or atoms known to be present must then conform in their mutual arrangement to the requirements of these elements of symmetry. At the time when the theory of space groups was first worked out by Fedorov, Schoenflies, and Barlow there must have seemed little possibility of assigning any actual crystal to its correct space group with certainty, except in very special cases. But with the advent of X-ray analysis the determination of space group has become practically a matter of routine, although difficulties and ambiguities exist in certain cases.

For several years following the discovery of the diffraction of X rays by crystals in 1912, very little use was made in England of the existing mathematical theory of crystallography. Attention was naturally focused at first on the more simple structures which could be worked out fully from first principles, and these simple structures were used in turn to analyse the X radiations produced from different sources. The importance of space-group theory, however, was realized in 1915 by Schoenflies[94], Fedorov[38] and Nishikawa[71]. Between 1918 and 1922 the analytical results of the theory were worked out and applied to X-ray analysis by Niggli[70] and Wyckoff[106]. With the application of X-ray methods to the more complex molecular and organic crystals the relation between crystal and molecular symmetry became a matter of great importance, and pioneer studies in this direction were made by

Shearer[93]. These results were worked out more fully and correlated with space-group theory in an important paper by Astbury and Yardley in 1924[2]. They classified in a simple way the sets of X-ray reflections which disappear in each space group (called "missing spectra" or "halvings" because the spacings of the corresponding planes appear to be halved), and they gave rules whereby the space groups could be differentiated on this basis. Shearer's rules relating the number of molecules in the unit cell to the symmetry possessed by the molecule itself were also worked out fully for every different space group.

These advances gave to research workers a method of applying the results of space-group theory to the X-ray analysis of crystal structure. If the unit cell of the structure contained only a few atoms, these could often be assigned directly to their correct positions, by the analytical results of Wyckoff, or all the possible alternative positions clearly indicated. In the majority, however, and especially in the complex molecular structures, large groups of atoms may be present which do not coincide with any symmetry element of the crystal. For the ultimate analysis of such structures we depend upon a study of the *intensities* of the various orders of the X-ray spectra, and no entirely systematic method of attack can be given. Such structures are considered later in this article. But even in these cases the initial step of determining the number of atoms in the asymmetric unit of the structure and the way in which these units are arranged in accordance with the symmetry of the crystal can be made systematically by the application of space-group theory. It is true that the mutual arrangement of these elementary groups of atoms in the crystal can often be determined correctly from first principles; and indeed there are many examples of this in the early literature. But the systematic application of space-group theory has the great advantage of presenting immediately all the mathematically possible alternative arrangements, some of which would otherwise very likely be overlooked.

Since 1924 there have been many further contributions to the geometrical side of crystallography and its application to X-ray analysis by Hermann[50], Mauguin[68] and others, much of it directed towards unification of the different systems of description. It is unnecessary to deal with this here owing to the recent publication of a number of accounts of present-day methods. New editions of standard works by the Braggs[14], Ewald[36] and by Wyckoff[107] have appeared, while the *International Tables for the Determination of Crystal Structure*[53] give a compact account of modern methods and nomenclature. More recently, the important step of summing the structure factor and electron density formulae (see §3) over the equivalent points in all the space groups has been carried out in great detail by Lonsdale[66].

Exceptional and random structures. In recent years a number of interesting structures have been discovered in which there appears a certain conflict between chemical and crystallographic evidence. We have seen that the systematic application of space-group theory leads to certain rules, that there are certain *equivalent positions* in the unit cell, whose coordinates have been tabulated by Wyckoff, and these should be occupied by identical atoms. In some structures it has been found, however, that a smaller number of atoms may be distributed through the specified

positions with a certain degree of randomness, an example being $\gamma - Al_2O_3$[42]. Again, in other structures, such as the spinels and lithium ferrite[81], two sets of atoms may be distributed indifferently between two sets of equivalent positions. The size of the ion rather than its chemical character is the dominating influence.

A discussion of such matters, which are analogous to the distributions found in metals and alloys, upon which a large amount of work has been done in recent years in connexion with the order-disorder transformation theory, is beyond the scope of this article. But they must be mentioned here because of their fundamental importance, and because they show that the straightforward application of the space-group rules may not in every case give the correct solution. There is not so much application to the molecular structures with which we are concerned, but reference should be made to the work of Hendricks[49] on the statistical centre of symmetry reported in *p*-chlorobromobenzene, and of Pauling[76] on ammonium oxyfluomolybdate.

Physical theory and X-ray methods. After Laue's discovery of X-ray diffraction in 1912[41], the first crystals to be completely analysed by the new method were NaCl and KCl by W. L. Bragg in 1913[15]. Meanwhile the method of the ionization spectrometer with monochromatic rays had been developed by W. H. Bragg[13] and was applied in his analysis of the diamond. Very soon a large number of other structures had yielded to this powerful new method of analysis, and descriptions of these will be found in the Braggs' well-known books[14].

Although the elucidation of these early structures depended essentially on the measurement of the relative intensities of the different orders of the reflected X-ray spectra, they did not require the use of absolute measurements, which involve the ratio of the intensity of the primary incident beam to that of the various reflected beams. Such absolute measurements introduce problems of great difficulty, both on the experimental side and in the calculation of the results, but they are of vital importance in the analysis of complex structures, as will appear later in this article. It is a most remarkable fact that early in 1914 not only were methods of defining and measuring the absolute intensities evolved by W. H. Bragg[9], but a complete theoretical treatment of the intensity of X-ray reflection was given by Darwin[28]. In these papers he calculated the two standard intensity formulae, for the mosaic and the perfect crystal, upon which all subsequent work in this field has been based.

The following year saw another important development, and this formed the subject of W. H. Bragg's Bakerian lecture to the Royal Society[10]. In it he developed for the first time the idea of the application of Fourier series to the problem of X-ray analysis. The existence of the crystal means a certain regularity of structure; some unit of pattern which can scatter X rays must be repeated in a periodic manner over a certain range. Any such periodic distribution of the density of the scattering matter may therefore be analysed by Fourier's method into a series of harmonic terms, and it can be shown that each X-ray reflection corresponds to one of these component sinusoidal distributions of density in the medium. The method of treatment was based on the analogy with the optical methods of Abbe and A. B. Porter, the former dealing with the case of image formation in the microscope.

Bragg realized that the coefficients of the Fourier series which represents the periodic variation of density in the medium should be proportional to the intensities of the corresponding X-ray reflections, but the method was not put into a usable quantitative form at this time owing to the many difficult and uncertain factors affecting the intensities of X-ray reflections.

The methods of using X rays in the study of crystalline matter were greatly extended in the next few years by a number of workers. The well-known powder method of Debye and Scherrer[31] and Hull[52] brought a great number of new materials within the range of X-ray analysis. The rotating crystal photographic method came into general use through the work of Schiebold and others, while Polanyi showed that many fibre structures also gave X-ray diagrams of this type[80]. The analysis of such rotation photographs, however, is a somewhat difficult problem, but reliable methods were introduced and the interpretation simplified by Bernal's treatment in 1926[6] based on Ewald's concept of the reciprocal lattice. A further important method of recording and indexing X-ray reflections photographically was introduced by Weissenberg[102] in 1924, by the use of cameras employing moving films. The original Laue method with a stationary crystal and so-called white radiation was meanwhile greatly developed and used with success in the analysis of a number of structures by Wyckoff[107], Pauling[75], Schiebold[93] and others.

During this time the theory of the intensity of X-ray crystal reflections and the problem of absolute measurements were extensively studied. A very full and independent treatment was given by Ewald[35] who arrived at the same formulae for the intensity of reflection as Darwin had previously given. In 1917 Compton[24] made absolute measurements and showed that the theoretical formulae were essentially correct, and was able to make an approximate calculation of the electron distribution in the sodium and chlorine ions. A very thorough and accurate determination of the absolute intensities in sodium chloride was made in 1921 by Bragg, James and Bosanquet[19].

These great developments in theory and technique had two immediate but far-reaching consequences for the science of X-ray crystallography. In the first place, the sound theoretical basis now established for the interpretation of intensities of X-ray reflections made it possible for the Fourier method of analysis to be put into a usable quantitative form. An early application of the method to long-chain structures was made by Shearer[96]. It was then rediscovered in an interesting way by Duane[33] in 1925 when he applied Epstein and Ehrenfest's quantum treatment of diffraction[34] to the determination of the distribution of diffracting power in a crystal. Duane's results were used by Havighurst[46] to calculate the distribution of electrons in a number of simple crystals. But perhaps the most important step was made by Compton[25], who derived the Fourier series expression from first principles, introducing the quantitative expression for the structure factor, and showing exactly how the coefficients are related to the X-ray intensities. He used the series in several different forms to calculate the electron distribution in a number of atoms. In 1929, W. L. Bragg[16] extended the method further and made use of a double

Fourier series for the first time, applying the results to his analysis of the diopside crystal. This elegant method gives a picture of the structure as seen in projection along certain crystallographic directions. In all this work, however, it must be emphasized that the Fourier series method does not lead to a direct solution of the structure except in special cases. In general, the phase constants of the terms must be ascertained by other methods; but still, the expression of the structure by means of a Fourier synthesis usually leads to a valuable refinement of the parameters involved, and gives a complete description of all the information that can be obtained from the available X-ray data.

In the second place it now became possible to apply the full power of the X-ray method to the unravelling of complex structures. Many experimental difficulties connected with the degree of perfection of the crystal specimens still remained, but these were attacked vigorously by W. L. Bragg and his co-workers in Manchester, and in a few years many structures of great complexity were completely solved. Results of the greatest importance to inorganic chemistry, and the physics of the solid state generally, were obtained. The structures of the complex silicates[17] are perhaps the best-known results obtained during this period (1924–30). The solution of such complicated structures could only be obtained by making the fullest possible use of absolute measurements of the intensities of the X-ray reflections. An excellent account of the methods involved has been given in a paper by Bragg and West[21] on "A technique for the determination of crystal structures with many parameters". Since then many still more complex structures have been successfully analysed, a recent example being $H_9PW_{12}O_4 . 29H_2O$[8].

In spite of these great advances very little progress was made until 1929 with the exact study of organic structures in the solid state. The difficulties which had prevented progress in this field were due to: (1) the complexity of the molecular structures which are crystalline at ordinary temperatures. Most of the simple organic compounds are liquids or gases; (2) the low symmetry of most of the crystals; (3) the fact that the compounds are composed almost entirely of H, C, N and O, the first having no appreciable influence on the scattering of X rays, while the other three are almost indistinguishable: the existing methods of analysis had depended to a large extent on the difference in scattering power of the elements composing the structure; (4) frequently only extremely small single crystals of organic compounds can be obtained, and with these it was impossible to apply the methods of measuring reflections, from extended faces, etc., which had already been developed. Consequently, only a few structures of unusual simplicity or high symmetry had been analysed, notable examples being certain long-chain hydrocarbons by Muller[69] and Shearer[96] and hexamethylene tetramine by Dickinson and Raymond[32]. The early measurements of W. H. Bragg[12] on organic compounds had shown that molecules existed in the crystal, and that these had dimensions in good agreement with calculations based on the interatomic distances in diamond and graphite[5] combined with the structural formulae of organic chemistry. These and other investigations gave reliable data for the space group and the symmetry of the molecule in the solid state, but did not lead to precise determinations of atomic positions.

A notable advance was made in 1929 when Lonsdale[65] determined the full structure of hexamethylbenzene. The benzene ring and the carbon atoms of the six substituent methyl groups were shown to lie accurately in one plane, and the dimensions of the ring corresponded with those of the carbon rings in graphite. The crystal had low symmetry, but the analysis was simplified by the fact that the unit cell contained only one molecule, and all the atoms were found to lie in one of the crystallographic planes.

As a consequence of this analysis it appeared probable that approximately planar models of graphite dimensions might apply to the majority of aromatic compounds. But if this were so, preliminary measurements showed that in these other compounds the atoms could not possibly all lie in any one crystallographic plane. The structures were evidently complicated, and further work was needed to overcome the difficulties mentioned under (1) to (4) above.

Methods for accurately recording and measuring the reflections from minute single crystals were required. Much preliminary work on the intensities of reflections from small crystals when completely bathed in a beam of X radiation was necessary. The method was the only one which it was possible to apply in the study of small organic crystals, and a thorough investigation of the absolute values of the reflections from small anthracene crystals by this means was undertaken by Robinson[91] in conjunction with a complete structural analysis by the writer[84] in 1932. The proportionality between the mass and the integrated intensity for small crystals up to about 0·25 mg. in weight was established, and questions of extinction and absorption were examined. Better methods for amplifying and recording minute electrical currents from the ionization spectrometer were an important factor in this work[105].

These methods were important for the initial determination of the scale of the absolute values of the reflections from small crystals, but were unsuited for the extended surveys required for complete structural analysis. The intensities of several hundred reflections were often necessary owing to the complexity of the structures. In this connexion the development of the photographic method for recording and measuring integrated intensities was of the greatest importance. An early form of integrating photometer was developed by Astbury[1] making use of the low penetrating power of α rays, but an improved photoelectric instrument was later produced by Robinson[92] and Dawton[29], while factors necessary for the correction of photographically recorded intensities were calculated by Cox and Shaw[27]. The extension of the photographic method to the determination of absolute values for the intensities was made by means of a two-crystal spectrometer (Robertson[85]) which enabled the reflections from an unknown crystal and a small standard crystal to be recorded side by side on the same moving film. The use of moving-film methods in such work is of the greatest importance in reducing the background surrounding the X-ray reflections, and so increasing the accuracy of the measurements.

This new technique has been applied from 1933 onwards in the accurate determination of a considerable number of complex molecular structures[82,83]. The planar configuration for aromatic compounds in general has been established, and the

accurate determination of bond lengths and valency angles in these and other molecular types is becoming an important connecting link between the empirical facts of chemistry and general theoretical developments. The method of double Fourier synthesis has been largely employed to refine the X-ray results by means of successive approximation, and the extensive numerical work involved has been simplified and put in manageable form by the methods of Beevers and Lipson[3, 63] and the writer[88].

We have seen that recourse is often necessary, especially in the preliminary stages of analysis, to other methods for indicating the probable arrangement of the atoms. Auxiliary methods of analysis of great importance have been developed in recent years by Krishnan and Banerjee[57], Lonsdale[67], and others[104], which depend on the optical and magnetic properties of the crystals. Another development of great interest (compare §4) is the direct Fourier series method of Patterson[72, 73].

§ 3. PHYSICAL THEORY

Definition of the structure factor, $F(hkl)$. In any study of crystal structure by means of X rays, two kinds of measurement can be applied to the various diffracted beams. These refer to their position and to their intensity. The position, denoted by the glancing angle θ, gives a measure of the spacing d of the corresponding crystal plane in terms of the wave-length λ and the order of the diffraction n, according to the Bragg relation

$$n\lambda = 2d \sin \theta \qquad \text{.....(1).}$$

By measuring the spacings of all the crystal planes, the size of the unit cell can be obtained, and the geometry of the crystal determined.

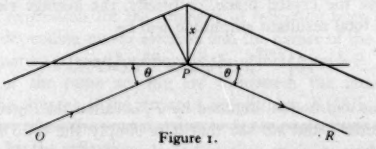


Figure 1.

In any study of the nature of the fundamental unit of pattern in a crystal, however, the relative intensities of the diffracted X-ray beams are of the greatest importance. These can be studied quite adequately by classical methods. Let us consider some distribution of scattering matter about the lattice point P figure 1. This scattering matter may consist of a single atom, a molecule, or a group of molecules—in general some distribution of electrons which is capable of scattering X rays. It is convenient to consider this portion of matter as representing the contents of the unit cell of the crystal, that is, the smallest unit which is repeated identically in every direction. An infinite series of possible crystal planes, of smaller and smaller spacing, intersect this unit, and of these we have selected one at random.

Let an X-ray beam be incident upon this plane at an angle θ. Under its influence all the electrons radiate waves, and we shall consider the intensity of the resultant

scattered beam in the direction R, also inclined at an angle θ to the crystal plane. In this direction the scattered beams from all the identical unit cells reinforce each other, in accordance with the interference relation (1), and finally build up sufficient intensity for experimental observation.

Consider first the wave scattered by a single electron situated at P. If the incident wave of amplitude A has its electric vector perpendicular to the plane containing OPR, then the amplitude A of the wave scattered by this electron can be calculated by classical electromagnetic theory[25] and is given by

$$\bar{A} = A \frac{e^2}{mrc^2} \qquad \ldots\ldots(2),$$

at a distance r from the scattering electron. We do not require to make use of this expression at present, but it is the basis of the absolute intensity formulae given in the next section.

Now the beam scattered by the whole unit is compounded from the waves radiated by all the electrons in this unit. Thus, if all these electrons lie exactly on the crystal plane, the component waves will all be in phase, and the resultant intensity will be a maximum. If the unit contains z electrons the amplitude of the resultant wave will be simply $\bar{A}z$. But if an electron is at a height x above the crystal plane, the path of the ray through this electron is increased by $2x \sin\theta$, and its contribution to the resultant amplitude is decreased by a factor

$$\cos 2\pi\,(2x \sin\theta/\lambda) = \cos 2\pi\, nx/d$$

by equation (1). If we now consider a symmetrical distribution about the crystal plane, say an atom of radius $a/2$, and if $\rho\,(x)$ is the probability of finding an electron at a height x above the crystal plane, or briefly, the average electron density at height x, then the total resultant amplitude will be

$$\bar{A} \int_{-a/2}^{+a/2} \rho\,(x) \cos 2\pi \frac{nx}{d}\, dx = F\bar{A}.$$

The integral in this expression, denoted by F, is called the "structure factor" or "structure amplitude", and we see that it is simply the ratio of the amplitude actually received from a given unit cell to that which would be received if the contents of the unit cell were replaced by a single electron.

In the general case the distribution of electrons in the unit cell will have no symmetry, and we have to express the structure factor for any plane, defined by the Miller indices (hkl) which express the reciprocals of the intercepts which the plane makes with the axes abc of the unit cell. In the simple one-dimensional case treated above, let us take the x direction, normal to the plane, as being the a axis, and let the plane be defined by the Miller indices ($h$oo). The fundamental spacing d can now be replaced by the axial length a, and the order of the reflection n by the index h. The path-difference of the ray through the electron at height x compared with one through the origin is then

$$2x \sin\theta = x \frac{n\lambda}{d} = \frac{hx\lambda}{a}.$$

The phase change involved is thus $2\pi\, hx/a$, and in general, when an electron at (xyz) contributes to the plane (hkl) the total phase change will be

$$2\pi\,(hx/a+ky/b+lz/c).$$

If $\rho\,(xyz)$ represents the density at the point (xyz), the resultant wave for the whole distribution, referred to an origin at (000), is

$$\bar{A}\int_0^a\int_0^b\int_0^c \rho\,(xyz)\,exp\;2\pi i\,(hx/a+ky/b+lz/c)\,\frac{V}{abc}\,dx\,dy\,dz=\bar{A}F\,(hkl)\;\ldots(3).$$

$F\,(hkl)$ is the structure factor for the general plane (hkl). The factor V/abc is necessary to complete the element of volume because the directions of the axes a, b, c depend on the geometry of the crystal, and may not always be at right angles to each other.

Relation of F (hkl) to intensity of reflection. It is obvious that the factor $F\,(hkl)$ defined by equation (3) expresses all that we can find out about the distribution of electrons in the ultimate units of the crystal through the medium of intensity observations. Methods of calculating the electron distribution from a knowledge of the F factors are given in the next section, but we must first enquire how these factors are related to the intensities of the reflections actually given by crystals. This matter involves calculations of considerable difficulty, in which the state of perfection of the crystal specimens plays an important part, and only a brief summary of the results can be given here. The first calculations were made by Darwin[28] and by Ewald[35], and simple treatments have been given by Compton[25] and by W. L. Bragg[14].

In the first place we see that $F\,(hkl)$ is a function of the structure only, and will in general be quite different for every different plane of reflection. The other factors entering into the expression for the intensity of reflection from crystals, however, will be constants depending on the electron and the nature of the radiation, such as the factor $\bar{A}$ in equation (2), and functions of the glancing angle θ. If reflections from different planes of the same spacing are compared, the intensities should be proportional to the squares of the structure amplitudes, $F\,(hkl)$. Actually, a qualification is necessary when the crystal structure is too perfect and the incident beam suffers extinction by the reflection.

The X-ray reflection from a crystal does not usually occur very sharply, but takes place over a small angular range in the neighbourhood of the position θ defined by equation (1). A satisfactory method of measuring the total energy reflected, called the "integrated reflection", was devised at an early date by W. H. Bragg[9]. If the crystal is turned through the reflecting position with angular velocity ω, the total energy E received in the ionization spectrometer or on the photographic plate will be inversely proportional to ω. If I_0 is the energy of the incident beam (energy per sec. per cm^2 at the crystal) the integrated reflection is measured by $E\omega/I_0$, a ratio which is characteristic of a given crystal plane, and which has been shown to be very nearly proportional to the volume of the crystal when certain very small specimens are employed[91].

In the case of a crystal element of volume δV, so small that its absorption of the

X-ray beam is not appreciable, the following expression for the integrated reflection can be calculated:

$$\frac{E\omega}{I_0} = \left[N \frac{e^2}{mc^2} F(hkl)\right]^2 \lambda^3 \frac{1+\cos^2 2\theta}{2 \sin \theta} \delta V \qquad \ldots\ldots(4).$$

In this expression N is the number of unit cells per unit volume of the crystal, or the reciprocal of the volume V of the unit cell. The factor e^2/mc^2 comes from equation (2), while $(1+\cos^2 2\theta)$ allows for the incident X-ray beam being unpolarized, the usual state when characteristic radiation is employed.

It can be shown that equation (4) should only be valid for extremely small, almost submicroscopic crystals. This condition may be realized by the use of a fine powder, but in practice it is found that equation (4) can be applied, with modifications allowing for the absorption of the X-ray beam, to certain large single crystals, such as rock salt in the experiments of A. H. Compton[24] and of W. L. Bragg, James and Bosanquet[19]. It must be inferred, as pointed out by Darwin[28], that such crystals consist of a conglomerate, or a *mosaic* of small blocks, "each block a perfect crystal, but the adjacent blocks not accurately fitted together". Other types of crystal imperfection can be imagined, but the idea of small blocks imperfectly aligned is more suitable for mathematical treatment. Moreover, if the crystal were a perfect structure throughout, the X-ray reflections should take place over a very narrow range, amounting to only a few seconds of arc, whereas in crystals of the rock-salt type the reflections are found to take place over a considerable angular range.

There are certain kinds of crystal, however, such as freshly cleaved calcite and some types of diamond, in which the X-ray reflection does take place very sharply over a narrow range of angle, and in these cases formula (4) cannot be applied. Such crystals approximate to the perfect type, and the theory of reflection for the perfect crystal has also been given by Darwin[28], who has shown that such crystals should reflect all the X rays incident upon them over a certain narrow range of glancing angle. In this case the integrated reflection is given by

$$\frac{E\omega}{I} = \frac{8}{3\pi} N \frac{e^2}{mc^2} F(hkl) \lambda^2 \frac{1+|\cos 2\theta|}{2 \sin 2\theta} \qquad \ldots\ldots(5),$$

I being the total intensity of the incident beam at the crystal face. It is to be noted that the structure factor $F(hkl)$ enters the expression as a first power only, in contrast to equation (4).

The mosaic crystal and the perfect crystal are two ideal types, for which an accurate mathematical treatment of the intensity of the X-ray reflections can be given. The crystals actually encountered in nature vary between these types, although they usually approximate more closely to the mosaic than to the perfect structure. With strong reflections and particularly when the crystal approximates to the perfect type, the extinction of the incident X-ray beam by reflection from the surface layers is much more important than ordinary absorption in the crystal. Two kinds of extinction are encountered. The effect is called "primary extinction" when the blocks of the mosaic are too large for equation (4) to be applicable. As there is no

method of measuring the size of the mosaic blocks, little can be done to correct for this effect. What is called "secondary extinction" occurs even in a good mosaic crystal, and is important for the very strong reflections. The lower part of the crystal is partially screened by the upper part reflecting away an appreciable amount of the radiation; the effective absorption coefficient of the crystal is thus increased when it is set at the reflecting angle. Experimental methods for detecting and dealing with this effect have been devised. The whole subject is one of great difficulty, and reference must be made to the works of Bragg, Darwin and James[18], Bragg and West[21] and James[54] for details. A useful summary of the various modifications of the intensity formulae required for different experimental conditions will be found in the *International Tables for the Determination of Crystal Structure*[53].

We have passed over the important subject of extinction somewhat superficially because these corrections do not play a large part in the technique which has been devised for dealing with the organic molecular crystals specially described in this article. This technique depends essentially upon the use of extremely small specimens (which are frequently the only ones available) combined with photographic methods or specially sensitive ionization spectrometers for recording the reflections, the small specimens being completely immersed in the incident X-ray beam. A number of factors contribute to the success of the method. It is found that the crystals generally belong to the mosaic type. The absorption of the X-ray beam is small, the component elements usually being H, C, N and O. Very strong X-ray reflections are not common because the structure amplitude is usually small owing to the large number of unsymmetrically arranged atoms in the unit cells. Because of these factors it is found that when allowance is made for the ordinary calculated absorption coefficient, formula (4) can be applied directly when the size of the crystal specimen is about 10^{-4} cm.^{3} or its weight is 0·1 to 0·2 mg. The specimens are usually cut, or grown in needle form, to a length of about 1 mm., which is set perpendicular to the direction of the X-ray beam, so that the path of the beam through the crystal is only about 0·2 or 0·3 mm. If there are specially strong X-ray reflections, due to the atoms all lying in some simple plane, or if many heavy atoms are present, allowance may still have to be made for secondary extinction effects. For details reference should be made to a critical study of small anthracene crystals by Robinson[91].

The principal aim of the analyses described in this article has been the accurate determination of the atomic positions in complex structures. The intensity measurements made by the above procedure can determine the values of $F(hkl)$ with amply sufficient accuracy for this purpose. But if a really critical survey of the electron distribution in the outer portions of the atoms is required, then a more careful study of the stronger X-ray reflections might be necessary, together with a study of the effect on the Fourier series of the inclusion of further terms which might be obtained by the use of longer exposures and shorter wave-lengths.

Expression of electron density by Fourier series. We have seen in the preceding section that, although considerable difficulties are involved, the magnitude of the structure factor $F(hkl)$ can usually be determined in a satisfactory manner from

experimental measurements of the intensity of the corresponding X-ray reflection. Equation (3) gave a relation between $F(hkl)$ and the density of the scattering medium, $\rho(xyz)$, which may be written

$$F(hkl) = \frac{V}{abc}\int_0^a\int_0^b\int_0^c \rho(xyz)\exp 2\pi i(hx/a + ky/b + lz/c)\,dx\,dy\,dz \ldots\ldots(6).$$

It is important to note that $F(hkl)$ for any given plane is a complex quantity, characterized by an amplitude and a phase constant. It is obvious that the experimental measurements of intensity, while defining the amplitudes of the structure factors $F(hkl)$, can give no information regarding these relative phase relationships. The different reflections are measured at different times with the crystal in different positions, hence information regarding the phase relationships is necessarily lost in making the experiment. This limitation of the X-ray method of analysis was realized at an early date. The ambiguity which is thus introduced represents the fundamental difficulty in X-ray analysis, and in the general case the problem of determining $\rho(xyz)$ from intensity measurements cannot be solved. But when other sources of information about the structure are considered in the course of the analysis, such as the existence of well-defined atoms, then a solution often becomes possible. Before considering these matters we must express the density $\rho(xyz)$ in terms of the structure factor $F(hkl)$ in a manner suitable for calculation.

Although we assume no symmetry in the density-distribution, the nature of the crystal requires the structure to be essentially periodic over a range defined by the primitive translations of the unit cell, a, b, c, and hence $\rho(xyz)$ can properly be represented by means of a Fourier series

$$\rho(xyz) = \sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty} A_{pqr}\exp 2\pi i(px/a + qy/b + rz/c),$$

p, q, and r being integers and A_{pqr} the unknown coefficient of the general term. If we can calculate A_{pqr} it will be possible to evaluate the electron density at any point in the crystal. To do this we substitute this series for $\rho(xyz)$ in the expression for the structure factor given by (6) and obtain

$$F(hkl) = \frac{V}{abc}\int_0^a\int_0^b\int_0^c \left[\sum\sum\sum_{-\infty}^{+\infty} A_{pqr}\exp 2\pi i\left(\frac{px}{a} + \frac{qy}{b} + \frac{rz}{c}\right)\right] \exp 2\pi i\left(\frac{hx}{a} + \frac{ky}{b} + \frac{lz}{c}\right)dx\,dy\,dz.$$

On integrating, every term is zero except that for which $p = -h$, $q = -k$, and $r = -l$, which gives

$$F(hkl) = \frac{V}{abc}\int_0^a\int_0^b\int_0^c A_{-h-k-l}\,dx\,dy\,dz = VA_{-h-k-l}.$$

The Fourier series which represents the electron density at any point may thus be written

$$\rho(xyz) = \sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty}\frac{F(hkl)}{V}\exp -2\pi i(hx/a + ky/b + lz/c) \quad \ldots\ldots(7).$$

The zero term in the series, $F(000)/V$, is a constant given by (6)

$$F(000) = \frac{V}{abc}\int_0^a\int_0^b\int_0^c \rho(xyz)\,dx\,dy\,dz = Z \qquad \ldots\ldots(8),$$

where Z is the total number of electrons in the unit cell. If the phase change on scattering is the same for all parts of the unit cell $F(hkl)$ and $F(\bar{h}\bar{k}\bar{l})$ are conjugate, and therefore $|F(hkl)| = |F(\bar{h}\bar{k}\bar{l})|$, in accordance with Friedel's law[40], and the series may be written

$$\rho(xyz) = \sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty} \frac{|F(hkl)|}{V}\cos\left[2\pi hx/a + 2\pi ky/b + 2\pi lz/c - \alpha(hkl)\right] \ldots\ldots(9),$$

where $\alpha(hkl)$ represents the phase constant of the structure factor $F(hkl)$.

We see, therefore, that if the electron-density distribution in the crystal lattice is analysed into a series of harmonic terms corresponding to the various possible crystal planes, then the coefficients of these terms are given by the absolute values of the structure factors for the corresponding planes, divided by the volume of the unit cell. A very beautiful reciprocal relationship is thus involved. The resultant wave scattered by the unit cell of the crystal is compounded from the contributions of all the electrons in the cell. There is an infinite series of such resultant waves corresponding to all possible crystal planes, and these resultant waves can themselves be compounded by means of a Fourier series to give an expression of the electron distribution in the unit cell.

The method is most readily applied to those structures which contain a centre of symmetry, an element which fortunately occurs quite frequently in crystals. When this symmetry is present in the unit cell as a whole, it must apply to each of the component waves which define the structure in the series given by equation (9). Hence if we take the origin of the coordinates (xyz) at the centre of symmetry, then either a peak or a trough of each of these component waves must coincide with the origin. The phase angle $\alpha(hkl)$, which measures the displacement of the peak of the wave from the origin, must in this case be limited to the values 0 or π, and these two cases can be covered by making the signs of the coefficients, $F(hkl)/V$, either positive or negative. When a centre of symmetry is present, the series may thus be written

$$\rho(xyz) = \sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty} \pm\frac{F(hkl)}{V}\cos 2\pi\,(hx/a + ky/b + lz/c) \qquad \ldots\ldots(10).$$

An important case in practice, which gives a maximum amount of information about the structure with a manageable amount of numerical work, arises when we confine the series to two coordinates, and express the density per unit area in any *projection* of the structure by the series

$$\rho(xy) = \sum_{-\infty}^{+\infty}\sum_{-\infty}^{+\infty} \pm\frac{F(hko)}{A}\cos 2\pi\,(hx/a + ky/b) \qquad \ldots\ldots(11).$$

In this case A represents the area of the cell face upon which the projection is made. This form of analysis was first used by W. L. Bragg[16] on the structure of diopside, and more recently, with the development of photographic methods for recording intensities, it has been employed very extensively in the analysis of complex organic

structures. It is particularly useful because all the reflections (*hk*o) from which the coefficients are calculated belong to one zone in the crystal. Thus if the crystal is rotated about any given axis, the reflections from the corresponding zone all occur in the equatorial plane, and the relative intensities can be conveniently recorded in one operation on a moving-film camera of the Weissenberg type[102] or on the two-crystal moving-film spectrometer[85]. When these intensities are measured and the series (11) finally evaluated, the result gives a projection of the structure along the axis about which the crystal was rotated. By putting together the results of two or more projections of this kind, the complete structure can often be reconstructed. Some examples of such projections are given in the following pages.

The simple one-dimensional series (12) gives the density per unit length averaged in sheets parallel to any given plane, and the successive orders of the reflections from that plane, (*h*oo), give the coefficients in the series, d being the spacing of the fundamental plane

$$\rho(x) = \sum_{-\infty}^{+\infty} \pm \frac{F(h00)}{d} \cos 2\pi \frac{hx}{a} \qquad \ldots\ldots(12).$$

This form is useful in simple structures, but does not give much help with more complex distributions.

It should be noted that the total number of electrons in any region of the structure can be found by integrating the density as given by equations (9) to (12), over the region required. When the density is integrated over the whole unit cell, we obtain the total number of electrons, Z. Thus when the series (9) is integrated over the whole volume V, all the terms become zero except the constant term involving F (ooo) which becomes equal to Z.

Practical application of Fourier series. In representing any crystal structure by means of a Fourier series, it is important to remember that the series is only convergent because the coefficients of the higher terms decrease in magnitude. It is apparent from figure 1 that at larger glancing angles there is greater interference among the scattered rays. Hence the values of F and the coefficients in the series do become smaller for terms involving high indices. The amount of falling off, however, depends upon how much the electron distribution is concentrated on the crystal plane. In a structure consisting of single atoms, the distribution is very concentrated near the centres, if the atoms are at rest. But at ordinary temperatures the atoms have considerable thermal movements and the electron distribution is effectively much more diffuse, with the result that at large glancing angles the intensities of the diffracted beams fall off quickly.

These considerations are of importance when interpreting the results of a Fourier synthesis. In practice, we cannot go on measuring the coefficients of the series indefinitely—a limit is set by the experimental conditions, particularly by the wave-length of the radiation used. Thus the limiting glancing angle of 90° corresponds to a minimum spacing of $\lambda/2$ by equation (1). Now if the coefficients of the Fourier series are still large when it is terminated by the experimental conditions, serious imperfections will occur in the density distribution which it represents. These defects take the form of false detail and regions of negative density which have no

physical meaning. This matter has been studied by Havighurst[47] for single Fourier series, and by Bragg and West[22] for double Fourier series. The latter authors make an interesting comparison of the projection obtained by double Fourier series with the case of the image formed by an optical instrument with a circular stop, and show that the defects in the Fourier projection of a crystal are like those produced by optical diffraction. If θ_0 is the limiting glancing angle up to which the spectra are measured, they show that the numerical aperture in X-ray analysis must be defined as $2 \sin \theta_0$, and that detail cannot be distinguished unless it is on a coarser scale than $0{\cdot}6\ \lambda/2 \sin \theta_0$.

When intensity-measurements are carried out at room temperatures, the thermal movement of the atoms is often large enough to ensure that the limiting terms in the series are very small compared with the initial terms. This is usually the case with light organic crystals, and the series are so rapidly convergent that the question of false detail need not arise if the measurements are carried out over a sufficient range of intensity. (A range of from several thousand units to one is often necessary.) In other crystals, chiefly hard minerals and inorganic compounds, comparatively strong reflections are still found at the limiting angle. In these cases the use of an artificial temperature factor of the form $\exp(-B \sin^2 \theta)$ has been suggested[23], in order to remove false detail from the projections by ensuring that the limiting terms in the series are small. This makes the peaks in the projection, which represent atoms, more diffuse, and so part of the information which might be obtained is sacrificed, but false detail is removed and electron counts become more reliable.

It is obvious that the larger the number of terms that can be included in the series, the more perfect the representation will become. This raises considerable difficulties, however, with regard to the numerical calculation of the results, particularly when double series are employed. For example, in the two projections of the phthalocyanine molecules shown in figure 5, a total of 281 different terms were used in the series, which had to be evaluated at 1800 separate points over half the molecule, in each case. If no shortening devices were employed, the total number of terms which it would be necessary to evaluate and add for these two projections would thus be over 500,000. Fortunately, a number of methods have been evolved in the last few years, chiefly by Beevers and Lipson[3, 63] and the writer[88], which shorten this work considerably.

If we write K for the coefficients and θ_1 and θ_2 for $2\pi x/a$ and $2\pi y/b$ in the double series given by equation (11) and expand, the expression for the density becomes

$$\rho(xy) = \Sigma [\Sigma K \cos h\theta_1] \cos k\theta_2 - \Sigma [\Sigma K \sin h\theta_1] \sin k\theta_2 \quad \ldots\ldots(13).$$

The summations in the brackets may be carried out first, and the totals obtained become the coefficients for the final summation, which may be written simply

$$\rho(xy) = \Sigma A \cos k\theta_2 - \Sigma B \sin k\theta_2.$$

The number of terms in the final summation is greatly reduced by this method, but the number of different coefficients, A and B, is increased, and it becomes worth while to make out a list of sine and cosine factors of all the natural numbers likely to be required, at some small interval, say 6°. In the method of Beevers and Lipson

strips of numbers are printed on cards covering all two-figure coefficients, positive and negative, for sines and cosines, and all values of the index h from 0 to 20. About 4000 cards are thus required, and the synthesis then reduces to selecting sets of cards and adding up successive columns of numbers. In the other method[88], three-figure coefficients are employed. This would make the number of cards by the above process excessive, and so only one card is employed for each coefficient. Positive and negative numbers, sines and cosines, and variations in the value of the index h are then taken care of by a semi-mechanical sorting device, and the synthesis again reduces to adding sets of figures.

When the structure has no centre of symmetry, exactly the same numerical methods can be employed[87], if the variable phase constant α in equation (9) is first combined with the coefficients. The double series

$$\rho(xy) = \Sigma\Sigma\, K \cos(h\theta_1 + k\theta_2 - \alpha)$$

can be put in the form

$$\rho(xy) = \Sigma\,[\Sigma\,(K' \cos h\theta_1 + K'' \sin h\theta_1)] \cos k\theta_2 - \Sigma\,[\Sigma\,(K' \sin h\theta_1 - K'' \cos h\theta_1)] \sin k\theta_2 \quad \ldots\ldots(14),$$

where $K' = K \cos\alpha$ and $K'' = K \sin\alpha$; and the above processes then applied with very little extra labour.

When the Fourier series is developed for a given crystal, care is needed in noting the way in which the phase constant varies in the different quadrants of the reciprocal lattice, where the numerical values of the structure factor may be the same. This depends on the symmetry. For example, in the space group C_{2h}^5 ($P2_1/a$) the origin of coordinates may be taken on a centre of symmetry and then $F(okl) = F(okl)$ if k is even, but $F(okl) = -F(okl)$ if k is odd. Similarly, $F(hko) = F(\bar{h}ko)$ when $(h+k)$ is even but $F(hko) = -F(\bar{h}ko)$ when $(h+k)$ is odd. These relations follow from a consideration of the geometrical structure factor (§4). The important step of working out all such relations completely for all the space groups and summing the expression for the electron density over all the quadrants has recently been completed by Lonsdale[66].

§ 4. ANALYSIS OF STRUCTURE

General methods. We saw at the beginning of the last section that lack of knowledge of the relative phase constants of the reflections made it impossible to apply the Fourier series method directly to the solution of structures unless other sources of information or assumptions about the structure were employed. The first assumption generally made is that the structure is built up from atoms, the atom being a small region where the density of scattering matter is high. The number and variety of atoms in each unit cell, and consequently the number of electrons in each atom, can generally be determined by chemical analysis and density measurements.

Now the scattering power of an atom for X rays, or the *atomic structure factor*, is fairly accurately known. The problem which remains to be solved is that of finding the positions of the various atoms in the unit cell. This can be expressed by means of a *geometrical structure factor*, with three parameters for each atom. Various likely

positions for the atoms, as indicated by the chemical structure and physical properties of the crystal (magnetic susceptibilities, refractive indices, etc.), can then be tested until agreement is obtained between the calculated and the observed intensities of reflections. No general rule can be given for this process, which we may call *analysis by trial*, but when a good measure of agreement is obtained and the main outlines of the structure are clear, the calculated structure-factor phase constants for all the more important reflections may be combined with the experimentally determined amplitudes, and a Fourier synthesis carried out. The results will generally lead to more precise values for all the atomic coordinates, and very often the position of some atom, previously quite uncertain, becomes clear. From these revised results, the phase constants of the weaker reflections, previously omitted, can now be calculated with some certainty, and the process repeated with a more complete Fourier series. In this way, by a process of successive approximation, the true structure can be obtained. The method is particularly important when a large number of parameters are involved which it is quite impossible to vary independently, although some guiding principle, such as the chemical structural formula, may make the initial outlines of the structure clear. This important application of Fourier series to crystal analysis as a means of successive approximation to the true structure does not seem to be generally understood, as it has sometimes been stated that such representations by Fourier series are merely elegant statements of the results of analysis achieved by other methods.

When some important atoms in the structure are found to lie in *special positions* with respect to the symmetry elements, a much more direct analysis is often possible. Space-group criteria may show that a given atom can only occupy some particular position in the unit cell. The contribution of this atom to the structure factor for every plane can then be calculated, and by a process of substitution the phase constant of the whole structure factor can sometimes be decided. This is perhaps the most direct kind of analysis possible, because apart from the atoms in the special positions we do not require to assume the existence of discrete atoms in the remainder of the structure at all.

Another direct method of analysis has been developed in recent years by Patterson[73]. A different form of Fourier series is employed, which represents a weighted average distribution of density about any point in the crystal. The coefficients are proportional to the squares of the structure factors and are independent of the phase constants. The results can give valuable information about certain interatomic distances in the crystal.

In this section we consider these different subjects briefly in turn, and in the next section some examples of recent structure determinations are given which illustrate their application.

Atomic structure factors. Much work has been done in determining the distribution of electrons in different kinds of atoms. This was, in fact, the central problem of most of the earlier quantitative work in X-ray analysis. Simple structures like rock salt and diamond, containing only one or two atoms in the unit of pattern, can readily be analysed, and the Fourier series expressions of the last section used to

determine the electron distributions. A more useful form of the series for this purpose is that employed by Compton[25] and others, which expresses the radial distribution, $U(r)$, of the electrons in the atom, $U(r)\,dr$ being the number of electrons contained between the shells of radii r and $r+dr$. Then it can be shown that

$$U(r)=\frac{8\pi r}{d^2}\sum_1^\infty hF(h00)\sin\frac{2\pi hr}{d} \qquad \text{.....(15).}$$

As the atom is small and most of the electrons which are effective in scattering concentrated near the centre, each component term in the general expression for the electron distribution in equation (10) will correspond to a wave with a peak at the centre of the atom; that is, the terms will all be positive in sign, at least until the spacing of the planes becomes much smaller than the size of the atom. Thus the ambiguity of the phase constant does not vitiate the results, provided that the analysis is not carried too far.

Owing to the spherical symmetry of the atom (see, however, recent work by Brindley[23] and Helmholz[48]) the structure factor $F(hkl)$ will be independent of direction in the crystal when a single atom is concerned, and will depend only on the spacing of the plane, d, or on $\sin\theta/\lambda$. Such structure factors, characteristic of a given atom, are usually denoted by f, and the curve obtained by plotting f against $\sin\theta/\lambda$ is called the "f curve". The theoretical calculation of such f curves from the Bohr orbits was first carried out by Hartree[44], and later the atomic models of wave mechanics were used[45]. Other treatments have been given by Thomas[98] and Fermi[30] and by Pauling and Sherman[77]. James and Brindley[55] have recently calculated the numerical values of f for all the atoms by these various methods, and the figures can be found in the *International Tables for the Determination of Crystal Structure*[53]. To compare theory with experiment, allowance must be made for the thermal movement of the atoms in the crystal. Much work has been done on the temperature factor by Debye[30], Waller[99,100] and others[56]. The subject is a difficult one, but of great theoretical importance. The temperature factor, however, has not entered the expressions given in this article, because we are chiefly concerned with the experimentally observed electron distributions, which of course include the thermal displacements.

The geometrical structure factor: analysis by trial. As far as a single atom is concerned, we may now assume that the structure factor f is known, at least with sufficient accuracy for our purpose. As this factor represents the ratio of the amplitude scattered by the whole atom to that scattered by a single electron at the centre, its application reduces the problem of analysis to that of finding the position of a comparatively small number of unit scattering sources in the structure. If S is a factor which takes account of the phase differences due to the positions (xyz) of these sources (geometrical structure factor), then

$$S(hkl)=\Sigma\exp 2\pi i\,(hx/a+ky/b+lz/c) \qquad \text{.....(16),}$$

the summation, which is for the plane (hkl), being taken over the positions of the

centres of all the atoms of one kind in the unit cell. [Compare derivation of equation (3).] The complete structure factor is then given by

$$F(hkl) = S_1 f_1 + S_2 f_2 + \ldots \qquad \ldots\ldots(17),$$

the suffixes, 1, 2, ... denoting atoms of different kinds.

The first step in the analysis of any complex structure lies in determining the space group, that is, the nature and arrangement of the symmetry elements of the crystal. This can usually be done by observing what sets of reflections, if any, are systematically missing from the X-ray spectra[3], together with certain observations on the form and piezo-electric properties of the crystal[104]. The space-group data in turn define the coordinates of the equivalent points in the crystal. Thus, if a representative point in the structure, say an atom, has the coordinates $(x_1 y_1 z_1)$ the operation of the symmetry elements ensure that a definite number of similar atoms exist at other points in the unit cell of the crystal, and the coordinates of these points can be obtained from a study of the symmetry elements present. These are the *equivalent points*, and their coordinates have been systematically tabulated in recent works[106, 53].

The next step lies in summing the geometrical structure factor, equation (16), over the coordinates of the equivalent points. If A and B represent the real and imaginary parts of this factor then the amplitude and phase constant are given by

$$|S(hkl)| = (A^2 + B^2)^{\frac{1}{2}} \qquad \ldots\ldots(18),$$

$$\alpha(hkl) = \tan^{-1}\frac{B}{A} \qquad \ldots\ldots(19).$$

As an example, we may take the space group $C_{2h}^5 (P2_1/a)$, in the monoclinic system, which is simple because of low symmetry, and important because of its frequent occurrence. The symmetry elements are dyad screw axes parallel to the b axis, glide planes of symmetry parallel to the (010) plane with glides of $a/2$, and certain consequent centres of symmetry. When an origin is taken on one of these centres of symmetry, the coordinates of the equivalent points are given by

$$(x_1, y_1, z_1); (-x_1, -y_1, -z_1); (a/2 + x_1, b/2 - y_1, z_1); (a/2 - x_1, b/2 + y_1, -z_1) \qquad \ldots\ldots(20).$$

When the expression (14) is summed over these points, we obtain

$$\left.\begin{aligned} A &= 4\cos 2\pi\left(hx/a + lz/c + \frac{h+k}{4}\right)\cos 2\pi\left(ky/b - \frac{h+k}{4}\right) \\ B &= 0 \end{aligned}\right\} \qquad \ldots\ldots(21),$$

the imaginary part being equal to zero because the origin was taken at a centre of symmetry. The expression can be further reduced for particular combinations of the indices (hkl). Thus when $(h+k)$ is even

$$A = 4\cos 2\pi(hx/a + lz/c)\cos 2\pi\frac{ky}{b} \qquad \ldots\ldots(22),$$

and when $(h+k)$ is odd

$$A = -4\sin 2\pi(hx/a + lz/c)\sin 2\pi\frac{ky}{b} \qquad \ldots\ldots(23).$$

The last expression is zero when k is zero, or when h and l are zero. It follows that the (hol) reflections are absent when h is odd, and that the ($o\bar{k}o$) are absent when k is odd. These are, in fact, the missing spectra which characterize this space group.

In space groups of higher symmetry with a large number of equivalent points the summation of equation (16) and its subsequent reduction to useful forms for special combinations of the indices is a tedious operation, and one in which mistakes can easily be made. Recently, however, this work has been completed in a comprehensive manner for all the 230 different space groups by K. Lonsdale[66], and the results are directly applicable to crystal analysis.

In the analysis of any structure by trial, the reduced expressions such as equations (22) and (23) must now be summed over the coordinates of all the atoms in the asymmetric portion of the unit cell, the contributions from atoms of different kinds being first multiplied by their appropriate atomic structure factors, in accordance with equation (17). The numerical aspect of this work is obviously analogous to the process of Fourier synthesis—compare equations (22) and (23) with (17)—and similar computational methods can be employed, especially for structures with many atoms of a similar kind. Details of such methods have recently been given[87].

An interesting way of evaluating structure factors in the course of trial analysis has lately been described by W. L. Bragg and Lipson[20]. This is the use of contoured graphs of the reduced structure-factor expressions. In any projection of a structure two coordinates define the position of each atom. The contour lines give the value of the structure factor for an atom in terms of these coordinates. Although the number of space groups is 230, the number of plane groups characteristic of the possible two-dimensional patterns obtained in projection is only 17[74], and some of these are very simply related. It thus becomes feasible to prepare structure-factor graphs covering all possible projections and for all the simple combinations of the indices h and k. In testing a possible structure, the positions of the atoms are plotted on the graph, and their contributions to the structure factor can be immediately read off with considerable accuracy. The most important aspect of the method, however, lies in a certain element of directness which it introduces to the analyses in being able to see in advance the effect of shifting an atom in a given direction. In some places where the contour lines are close together, a small shift will have a big effect, while other positions are comparatively insensitive. The method is most valuable in dealing with structures of high symmetry where the analytical expressions are complicated, and which contain a relatively small number of atoms.

Atoms in special positions. We have already used the space group C_{2h}^5 ($P2_1/a$) to illustrate some general methods, and the coordinates of the equivalent points referred to a centre of symmetry as origin are given by equation (20). An atom put in the general position ($x_1y_1z_1$) will be found four times repeated, in accordance with these relations, in different parts of the unit cell. Now suppose that chemical analysis, the size of the unit cell, and the density of the crystal show that only two atoms of a certain kind are present in the unit cell. A consideration of equation (20)

will show that this can only happen if these two atoms coincide with centres of symmetry. Their possible coordinates are

$$(a)\ 000;\ \frac{a}{2}\frac{b}{2}0.\quad (b)\ 00\frac{c}{2};\ \frac{a}{2}\frac{b}{2}\frac{c}{2}.\quad (c)\ \frac{a}{2}00;\ 0\frac{b}{2}0.\quad (d)\ \frac{a}{2}0\frac{c}{2};\ 0\frac{b}{2}\frac{c}{2}\quad \text{.....(24).}$$

(Note that owing to periodicity over the range abc, $+a/2$ is equivalent to $-a/2$, etc.) If an origin is taken on one of these two atoms, the coordinates must be

$$(a)\ 0\,0\,0;\ a/2,\ b/2,\ 0.$$

If there are only two atoms in the cell altogether, the problem of position is completely solved, without recourse to intensity measurements or Fourier analysis. This occurs only in the most simple structures. But quite often in complicated structures it is found that atoms of a certain kind, owing to their restricted number, must occupy special positions, while the remaining atoms which build up the structure may be anywhere.

In other space groups the special positions may not be so definite as those given above. But in a given projection of a structure some of the atoms may be fixed or restricted to a few alternative positions. Furthermore, these few atoms so restricted are often heavy atoms which make a large contribution to the structure factor, while the remaining atoms in the general positions may be comparatively light, like oxygen or carbon, and so make much smaller contributions. By observing the absolute values of the structure factors for given planes, and knowing the contribution which must be made by the heavy atoms, the possible positions of the lighter atoms can often be restricted to certain narrow regions of the structure. Such methods are well illustrated in the analysis of diopside and other silicates by W. L. Bragg and West[21].

An excellent example, this time of a more direct determination, is provided by West's analysis of potassium dihydrogen phosphate[103]. The restricted positions of the potassium and phosphorus atoms govern the signs or phase constants of certain sets of reflections. A preliminary Fourier synthesis employing only these sets of reflections gives some indication of the positions of the oxygen atoms. These roughly determined positions then enable the phase constants of the other sets of reflections to be calculated. Repeated application of Fourier syntheses then leads to a complete and accurate determination of the structure.

The most direct pictures of complex structures, however, have been obtained by the study of certain isomorphous series of crystals, containing some variable element in a special position. This element can be substituted or exchanged for another of different scattering power, or perhaps removed altogether, without greatly disturbing the remainder of the structure. The effect of making such a substitution is to change the absolute values of the structure factors by a known amount but in a direction depending on the phase constant of the reflection. Hence by observing the directions of these changes a direct determination of the phase constants can be made, and a straightforward Fourier synthesis of the structure carried out. The best examples of structures determined in detail by this method are provided by an important series of organic compounds called the phthalocyanines.

and these are illustrated in the next section. The alums form another isomorphous series of crystals capable of analysis by this method, and an early application of single Fourier series to these structures was made by Cork [26]. More recently, a number of very beautiful determinations by the use of double Fourier series have been carried out by Lipson and Beevers [61, 62].

Direct F^2 Fourier series of Patterson [73, 101]. This method may be illustrated by considering a one-dimensional distribution of scattering matter, say along the a axis of a crystal.

The average density in planes parallel to (100) is given by

$$\rho(x) = \frac{1}{d} \sum_{-\infty}^{+\infty} F(h00) \exp 2\pi i \frac{hx}{a}$$

by equation (7), $F(h00)$ being equal to $F(\bar{h}00)$. Patterson then defines a quantity, $A(u)$, called the weighted average distribution of density about the point x by the relation

$$A(u) = \frac{1}{a} \int_0^a \rho(x)\, \rho(x+u)\, dx \qquad \ldots\ldots(25).$$

In this expression, $\rho(x+u)$ gives the distribution about x as a function of the parameter u, and this distribution is weighted by the amount of scattering matter between x and $x+dx$, given by $\rho(x)\, dx$. $\rho(x)$ and $\rho(x+u)$ are now expanded by the Fourier series given above, and the expression for the weighted average distribution becomes

$$A(u) = \frac{1}{d^2} \sum_{-\infty}^{+\infty} F^2(h00) \exp 2\pi i \frac{hu}{a} \qquad \ldots\ldots(26).$$

Now $A(u)$ depends upon the squares of the structure factors and is therefore independent of the signs or phase constants, and can be evaluated directly from the X-ray intensity-measurements. Its importance in crystal analysis arises because, by equation (25), it can only attain a large value when both $\rho(x)$ and $\rho(x+u)$ are large. Hence if the structure contains atoms at a distance u_1 apart, we may expect a maximum in the function $A(u)$ at this point.

In three dimensions

$$A(uvw) = \frac{1}{V} \int_0^a \int_0^b \int_0^c \rho(xyz)\, \rho(x+u, y+v, z+w)\, dx\, dy\, dz$$

$$= \frac{1}{V^2} \sum_{-\infty}^{+\infty} \sum_{-\infty}^{+\infty} \sum_{-\infty}^{+\infty} F^2(hkl) \exp 2\pi i\, (hu/a + kv/b + lw/c) \ldots\ldots(27).$$

$A(uvw)$ will be large only when there are high values in $\rho(xyz)$ both at (x, y, z) and $(x+u, y+v, z+w)$. Thus a peak in the function $A(uvw)$ at $(u_1v_1w_1)$ corresponds to an interatomic distance in the crystal defined by the vector whose components are $(u_1v_1w_1)$.

There remains the essential ambiguity which arises from the fact that it is not directly possible to decide to which pairs of atoms in the structure these vectors refer. When the number of atoms is large, the number of possible sets of interatomic distances is very great, and the method becomes complicated. But in some structures which contain only a few heavy atoms of large scattering power, the method

has been applied with success, in the form of two-dimensional series, by Patterson[73], Beevers and others[4, 37]. Patterson has also developed methods for improving the definition of the function, and for the removal of unnecessary peaks.

A valuable extension of the method has recently been made by Harker[43], in the use of three-dimensional series simplified by the symmetry properties of the crystal. For example, if the crystal has a twofold axis, an atom at (xyz) is accompanied by one at the equivalent point $(-x, y, -z)$, and the vector between these atoms has the components $(2x, 0, 2z)$. Hence there is a maximum in the function $A(uvw)$ at $(2u, 0, 2w)$ corresponding to this interatomic distance. To find the position of this peak, it is only necessary to evaluate $A(uvw)$ in the plane $v=0$. For the calculation, absolute values are unimportant, and the series may be written simply

$$A(u0w) = \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F^2(hkl) \cos 2\pi (hu/a + lw/c)$$

$$= \sum_{h=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} \left\{ \sum_{k=-\infty}^{+\infty} F^2(hkl) \right\} \cos 2\pi (hu/a + lw/c) \quad \ldots\ldots(28).$$

As the series is effectively only two-dimensional, the calculation is practicable. Good definition is obtained because all the measurable values of $F(hkl)$ are used, and because interatomic vectors not parallel to the plane $v=0$ are eliminated.

When the crystal has a plane of symmetry, the result is even more simple. The equivalent atoms are at (x, y, z) and $(x, -y, z)$, and the vector components are $(0, 2y, 0)$. The maximum in $A(uvw)$ lies on the b axis, and the series becomes one-dimensional:

$$A(0v0) = \sum_{k=-\infty}^{+\infty} \left\{ \sum_{h=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F^2(hkl) \right\} \cos 2\pi \frac{kv}{b} \quad \ldots\ldots(29).$$

Other forms of the series are discussed by Harker which take advantage of different types of symmetry elements which may be present.

The method, of course, leaves one or two coordinates undetermined. Although the ambiguities of the ordinary two-dimensional projection series are reduced, difficulties may still arise through atoms which are not related by the symmetry elements having by chance one or two coordinates alike.

§ 5. EXAMPLES OF RECENT ANALYSIS

Oxalic acid dihydrate. This is an example of a structure first determined by methods of trial, and later refined by the application of double Fourier series to two zones of structure factors. It illustrates the importance of such an application of the Fourier method, because although the structure is fairly simple, and can be determined with quite reasonable accuracy by the method of trial, yet the information which is of real importance to theoretical chemistry can only be obtained from the results of the Fourier synthesis.

Oxalic acid dihydrate, $C_2H_2O_4$, $2H_2O$ crystallizes in the monoclinic system, space group C_{2h}^5 $(P2_1/n)$, with $a=6{\cdot}12$, $b=3{\cdot}60$, $c=12{\cdot}03$ A., $\beta=106{\cdot}2°$, and there are two molecules of the above formula in the unit cell. Each molecule must there-

fore have a centre of symmetry, which coincides with one of the centres of symmetry in this space group. The total number of electrons in the unit cell, $F(000)$, is 132. The equivalent points in the unit cell are similar to those given in equation (20) with the addition of $c/2$ to the last two z coordinates. (This is merely due to a somewhat arbitrary choice of axes.) The asymmetric portion of the structure which has to be determined is therefore a quarter of the whole, or one carbon atom and two oxygen atoms in the acid molecule, together with a water molecule. The hydrogen atoms are as usual beyond the reach of X-ray analysis.

The trial analysis of this structure was carried out very carefully by Zachariasen[108]. There are some features which help to simplify the work. Thus, the very short b axis and the optical properties of the crystal (high negative birefringence) indicate that the molecules are probably nearly flat, and inclined at a fairly small angle to the (010) plane. When the intensities of the ($h0l$) reflections are examined in detail, there is a further important indication of an approximate symmetry in the structure about the [10$\bar{1}$] axis (or the long diagonal of the ac face).

Starting from this basis, extensive calculations of the structure factors by equations (16) and (17) for different positions of the atoms finally led to a good agreement with the observed values for one particular configuration. We need not give the coordinates here, but the dimensions of the molecule finally obtained are shown in figure 2, and the water molecules were placed about 2·6 A. away from the carboxyl oxygen atoms.

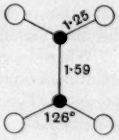


Figure 2. Oxalate group.

Now as 12 parameters are involved in this structure, for the four atoms in the asymmetric group, it is quite impossible to vary all these parameters independently in the course of the trial analysis. The result in figure 2 is rather a model of the molecule, which must be approximately correct. It must be fairly flat and symmetrical. The distances between the heavier atoms like 1 and 2 are probably correct to better than 0·1 A., whereas the error in placing the lighter carbon atoms may be greater. This model in a certain orientation explains the intensities of the X-ray spectra very well; but it is impossible to say, by trial analysis, whether some slightly different model in some slightly different orientation might not be a little better, because every possibility cannot be tested.

The agreements obtained with the observed structure factors, however, are good enough to determine the phase constants of these quantities (0 or π) down to very small values. A Fourier synthesis can thus be carried out, using the measured structure factors and the phase constants calculated from Zachariasen's coordinates. Now the great advantage of the application of Fourier series to the problem is that it presents a unique solution, making the best use of all the data. The accuracy depends simply on the accuracy of the intensity measurements and the completeness of the series employed.

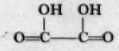
Two projections of this structure were made, employing double Fourier series in equation (11), in directions at right angles to each other, along the a and b crystal axes. The one which gives most direct information about the structure, along the

b axis, is reproduced in figure 3 (Robertson and Woodward[90]). The area covered is two unit cells, in order to show the surroundings of the molecules. $\rho(xz)$ is represented over this area by means of contour lines drawn at unit intervals (one electron line dotted) by graphical interpolation from the summation totals. The atoms appear reasonably circular, and there is no difficulty in distinguishing the higher peaks of the oxygen atoms from the two lower ones due to carbon atoms. Two coordinates for each atom can be measured directly from this map (originally drawn on a scale of 5 cm. to 1 A.), and the third can be determined from the other projection.

The dimensions of the molecule obtained from these results are shown in figure 4. The accuracy in placing the atoms is now estimated at from 0·01 to 0·04 A., depending on the degree of resolution of the different atoms. On the whole, the agreement with the previous trial model is quite as good as can be expected, but the increased precision and definiteness now attained are of great importance. The accuracy might be increased still further. From the new coordinates, the structure factors could be calculated again, and a few more phase constants of still weaker reflections determined. This would lead to a more complete series and still more accurate definition.

The results as they stand, however, are quite sufficient to enable certain important conclusions to be drawn regarding the structure of the molecule. The normal single-bond carbon-carbon distance is 1·54 A. (in diamond). The much smaller value of 1·43 A. in oxalic acid shows that this is not a true single bond, but that some alternative excited structure with a double bond between the carbon atoms must make a contribution to the normal state of the molecule. This is an example of resonance, a phenomenon which is being recognized as of extreme importance in organic chemistry[78]. Further evidence of this state is obtained from the fact that the atoms in the acid molecule are all found to lie in one plane, which is, however, inclined to the plane of the projection in figure 3.

The full application of such precise measurements of interatomic distance cannot be discussed here, but mention may be made of recent theoretical calculations, by Lennard-Jones[58] and Penney[79], of the carbon-carbon distances in butadiene $CH_2{=}CH{-}CH{=}CH_2$. The length predicted for the central bond is 1·43 A. As butadiene is a gas, precise measurements are difficult, but in oxalic acid there is a similar conjugation of the double bonds in the carboxyl groups



We might therefore expect a contraction of about the same order as that calculated for butadiene.

The positions of the water molecules shown in figure 3 indicate the presence of intermolecular bindings known as hydrogen and hydroxyl bonds. This is a type of association which can be clearly recognized from crystal structure measurements[7, 83].

The phthalocyanines[86, 89]. These structures provide a good example of analysis carried out directly from the X-ray intensity measurements. In deriving the diagram

Figure 3. Oxalic acid.

Reproduced by permission of the Chemical Society)

in the centre of figure 5, for example, we do not require to assume the existence of atoms at all, but merely some general distribution of electrons capable of scattering X rays.

The phthalocyanines are a series of complex organic compounds, extensively studied by Linstead and his collaborators[59], which are related in an interesting

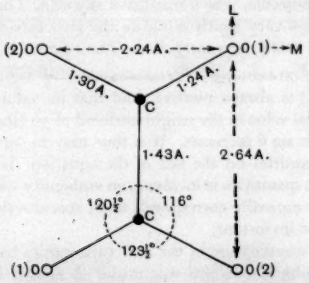


Figure 4. Dimensions of oxalic acid molecule.

manner to the natural porphyrins. They have the formula $C_{32}H_{16}N_8M$, where M is a metal atom. This metal atom can be removed, and the free compound has the formula $C_{32}H_{18}N_8$. This compound crystallizes in the monoclinic system, space group C_{2h}^5 ($P2_1/a$), $a=19{\cdot}85$, $b=4{\cdot}72$, $c=14{\cdot}8$ A.; $\beta=122{\cdot}2°$, with two molecules per unit cell. The data for a large number of the metal compounds are very closely similar[60], and they form an isomorphous series of crystals.

As in oxalic acid, the molecules must have centres of symmetry, coinciding with two of the crystal centres of symmetry. If we take an origin at one of these centres, the other must be at ($a/2$, $b/2$, 0) by equation (24). Now as there is only one metal atom in each molecule of the metal derivatives, this atom must occupy the symmetry centre. The positions of the metal atoms and their contributions to the structure amplitude are thus definitely fixed by the space group. By equations (22) and (23), this contribution is a maximum for all planes with $(h+k)$ even, and zero for all planes with $(h+k)$ odd.

To proceed with the analysis, we can first show that the unknown structures in two isomorphous members of the series are essentially the same, apart from the central metal atom. This can be done by comparing the absolute values of the structure factors for corresponding planes in the two members when $(h+k)$ is odd, i.e. when the metal atoms make no contribution. The measured values are then found to be very nearly the same in every case, and therefore the structures are also nearly the same. The only other necessary condition is one already discussed under atomic structure factors in the last section—that most of the electrons in the central metal atom are concentrated in a small region, which ensures that the contributions

of these atoms to the various structure amplitudes are always positive, for planes within the measurable limits.

The metal-free compound and the nickel derivative were chosen for detailed comparison. Now, for a given plane, with $(h+k)$ even, the observed structure amplitude from the nickel derivative is the algebraic sum of two terms, the contribution due to the nickel atom alone, which is always positive, and the contribution from the rest of the molecule, which may have any sign. This last contribution, as we have seen, is always very nearly equal to the structure factor from the corresponding plane in the metal-free compound. The relation involved is simply

$$F_{(\text{Ni derivative})} - F_{(\text{Free compound})} = 2F\,(\text{Ni}).$$

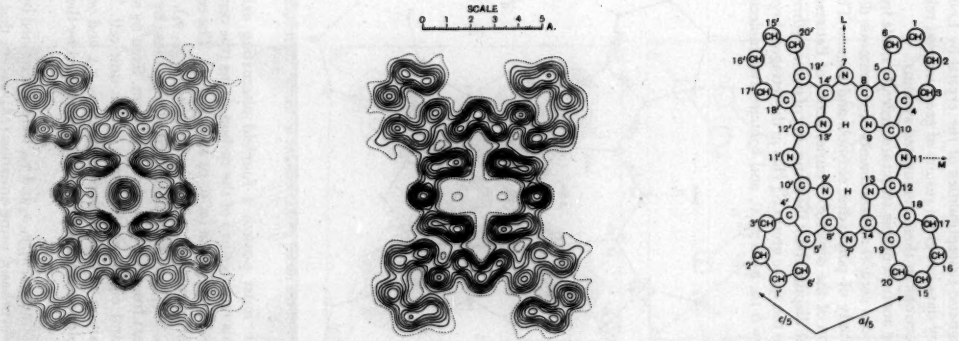
We know that $2F$ (Ni) is always positive, and that its value should decline continuously from an initial value in the neighbourhood of 56 (the number of electrons in two nickel atoms) as sin θ increases. It is thus easy to settle the correct sign of the two measured quantities on the left of the equation. In general, as a sum or difference of these two quantities is involved, an ambiguity will only arise when one of them is small. This normally corresponds to an absent reflection, or one so weak that its omission is not important.

From a numerical examination of the structure factors by this method a direct determination of the phase constants was made. A double Fourier synthesis was then carried out, by equation (11), for the (h0l) sets of structure factors in the two compounds, where only h even terms occur, by equation (23). The results gave the projections of the structures along the b crystal axes which are shown in figure 5, where $\rho\,(xz)$ is plotted by contours at unit intervals.

In these projections the electron distribution is clearly segregated into peaks, which there is no difficulty in identifying as due to carbon and nitrogen atoms. The latter rise to slightly higher values, and are found to contain about one more electron, when $\rho\,(xz)$ is integrated over their areas. The atoms are further clearly seen to be arranged in the 5- and 6-membered rings familiar to chemistry. On the central nickel atom the peak is actually about six times as high as on the carbon and nitrogen atoms, but only every fifth contour line has been drawn for this atom.

For the detailed working out of these diagrams reference must be made to the original papers[86, 89]. We note, however, that two coordinates for each atom can be determined directly from the maps, and the third can be obtained by various methods. The complete structure shows that all the atoms lie in one plane, and that this plane is inclined at about 45° to the planes of the projections given above. The molecule itself has almost tetragonal symmetry, and the dimensions for the nickel derivative are shown in figure 6.

The most interesting part of the structure is, perhaps, the inner nucleus consisting of a 16-membered carbon-nitrogen ring, to which the molecule appears to owe its great stability. This nucleus cannot be represented by a definite chemical formula, but displays single-bond—double-bond resonance similar to that found in benzene. There is no doubt that the stability of the natural porphyrins is due to a similar configuration, but in these structures the outer nitrogen atoms are replaced by CH groups.



(*Reproduced by permission of the Chemical Society*)

Figure 5. Electron density maps of the phthalocyanine molecules.

A study of the magnetic susceptibilities of the crystal by Lonsdale[67] in conjunction with this X-ray data has given results of great interest relating to the possible orbits of the π electrons in these conjugated systems.

Pentaerythritol. The analysis of this structure, recently made by Llewellyn, Cox and Goodwin[64], provides certain measurements of interatomic distance which are of importance in the study of polyhydric alcohols and sugars. The methods used illustrate the application of the direct F^2 Fourier series of Patterson.

There has been some dispute about the symmetry of pentaerythritol, $C(CH_2OH)_4$, but it is now assigned to the space group S_4^2 $(I\bar{4})$ in the tetragonal system. The body-centred unit cell has $a = 6{\cdot}10$, $c = 8{\cdot}73$ A., and contains two molecules of the above

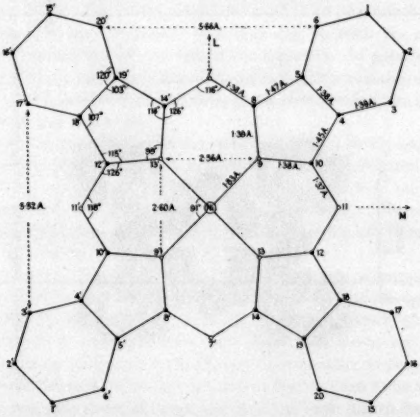


Figure 6. Dimensions of the nickel phthalocyanine molecule. o nitrogen; • carbon.

formula, each of which must possess a fourfold alternating axis of symmetry parallel to the c axis. This requires two of the carbon atoms (the central one in each molecule) to be in special positions at (0, 0, 0) and $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. The equivalent points in the unit cell are eight in number, with coordinates (x_1, y_1, z_1), $(-x_1, -y_1, z_1)$, $(y_1, -x_1, -z_1)$, $(-y_1, x_1, -z_1)$ and four more derived from these by the addition of $(a/2, b/2, c/2)$. Now apart from the central carbon atom, fixed by symmetry, there are only eight carbon and eight oxygen atoms in the unit cell. Thus it is only necessary to fix the position of one carbon atom (x_1, y_1, z_1), and one oxygen atom (x_2, y_2, z_2), when the whole structure is completely determined by the relations given above.

Four of these six parameters, x_1, y_1, x_2 and y_2, were determined by an application of the Patterson-Harker synthesis, using equation (28), except that in this case the

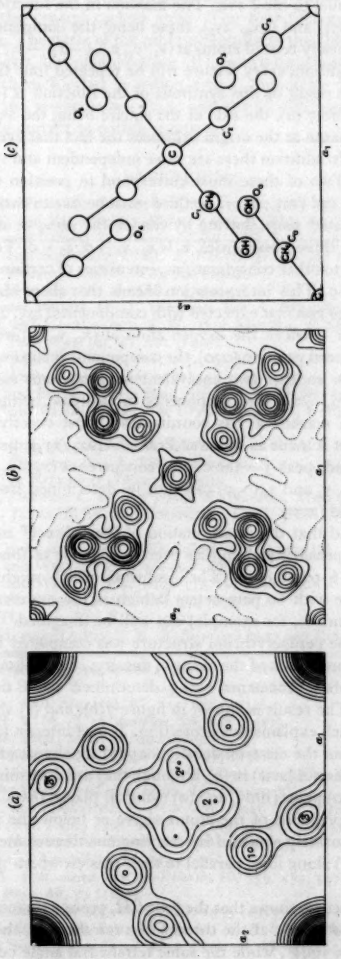


Figure 7.

symmetry axis is parallel to the c axis. The maxima in the function $A(uv0)$ should correspond to $(2x_1, 2y_1)$ and $(2x_2, 2y_2)$, these being the components of the vector between the symmetrically related atoms at (x_1, y_1, z_1), $(-x_1, -y_1, z_1)$, and (x_2, y_2, z_2), $(-x_2, -y_2, z_2)$. In addition, every feature will be repeated four times by the tetragonal symmetry. The result of this synthesis of the function $A(uv0)$ on the plane $w=0$ is shown in figure 7 (a), the side of the square being the axis $a=6{\cdot}10$ A.

The intense maximum at the origin expresses the fact that every atom is at zero distance from itself. In addition there are three independent and very distinct peaks marked 1, 2 and 3. Two of these must correspond in position to the vectors we require, viz. $(2x_1, 2y_1)$ and $(2x_2, 2y_2)$. The third must be due to two atoms, other than the symmetrically related pairs, having by chance the same or nearly the same z coordinate. There are three possibilities, $z_1 \fallingdotseq z_2$, $z_1 \sim 0$, $z_2 \sim 0$. To settle this point, recourse must be had to other considerations, intensities of certain spectra, etc.; and it is shown that $z_2 \sim 0$. This interpretation means that there should now be four independent peaks; the two first expected with components $(2x_1, 2y_1)$ and $(2x_2, 2y_2)$; one due to the vector between the oxygen atom at (x_2, y_2, z_2) with $z_2 \sim 0$ and the carbon atom in the special position (000), the components being (x_2, y_2); and one due to the vector between another two equivalent oxygen atoms each with $z_2 \sim 0$ at (x_2, y_2, z_2) and $(-y_2, x_2, -z_2)$, the components in this case being (x_2+y_2, y_2-x_2). Noting that the peak marked 2′ has coordinates almost exactly double those of peak 1, the assignment is made as follows. Peak $2' = (2x_2, 2y_2)$; peak $1 = (x_2, y_2)$ and $(2x_1, 2y_1)$ superimposed; peak $3' =$ the chance concurrence (x_2+y_2, y_2-x_2). The five coordinates x_1, x_2, y_1, y_2 and z_2 (~ 0) can thus be determined from a study of this contoured section of $A(uvw)$.

It will be observed that the interpretation of the "direct" method is not free from difficulties and possible ambiguities even in the case of this relatively simple structure containing 6 parameters. The confusion which might arise if it were applied to a structure with 60 parameters (which is the number involved in any one of the phthalocyanines, for example) can well be imagined.

The analysis of the pentaerythritol structure was completed by making a two-dimensional Fourier projection of the electron density, $\rho(xy)$, along the c axis, using equation (11), the phase constants being determined from the results of the Patterson synthesis. The result is shown in figure 7 (b), and (c) shows the arrangement of the atoms which explains this projection. It is of interest to note that the coordinates obtained from the electron density map (b) are almost identical with those derived from the function $A(uvw)$ in (a), although they are determined from different data, (b) using the ($hk0$) planes only, and (a) using all planes (hkl). The coordinates, z_1 and z_2, giving the positions of the atoms above or below the plane of the projection, were then accurately obtained by carrying out three-dimensional syntheses of $A(uvw)$ and $\rho(xyz)$ along lines parallel to the c axis on which the atoms are seen to lie in figure 7.

The complete structure shows that the four CH_2 groups surrounding the central carbon atom lie almost exactly at the corners of a tetrahedron, the angles differing by only about 2° from 109½°, while the same tetrahedral angle occurs between the

CH_2 groups and the OH groups, bringing these latter almost exactly back to the plane containing the central carbon atom. The four OH groups thus lie at the corners of a square. The parts of the other molecules in the corners of figure 7 (*c*) lie half-way up the *c* axis, but neighbouring OH groups on adjoining molecules at the same level approach each other to within 2·69 A., indicating hydroxyl bonds. The C—C distance in the molecule is 1·50 A., and the C—OH distance 1·46 A., values which are probably correct to within ± 0·03 A.

REFERENCES

(1) Astbury, W. T. *Proc. roy. Soc.* A, **115**, 640 (1927); **123**, 575 (1929).
(2) Astbury, W. T. and Yardley, K. *Philos. Trans.* A, **224**, 221 (1924).
(3) Beevers, C. A. and Lipson, H. *Nature, Lond.*, **137**, 825 (1936).
(4) Beevers, C. A. and Schwartz, C. M. *Z. Kristallogr.* **91**, 157 (1935).
(5) Bernal, J. D. *Proc. roy. Soc.* A, **106**, 749 (1924).
(6) Bernal, J. D. *Proc. roy. Soc.* A, **113**, 117 (1927).
(7) Bernal, J. D. and Megaw, H. D. *Proc. roy. Soc.* A, **151**, 384 (1935).
(8) Bradley, A. S. and Illingworth, J. W. *Proc. roy. Soc.* A, **157**, 113 (1936).
(9) Bragg, W. H. *Phil. Mag.* (6), **27**, 881 (1914).
(10) Bragg, W. H. *Philos. Trans.* A, **215**, 253 (1915).
(12) Bragg, W. H. *Proc. phys. Soc.* **34**, 33 (1921).
(13) Bragg, W. H. and Bragg, W. L. *Proc. roy. Soc.* A, **89**, 277 (1913).
(14) Bragg, W. H. and Bragg, W. L. *The Crystalline State*, **1** (1933). London: Bell and Sons.
(15) Bragg, W. L. *Proc. roy. Soc.* A, **89**, 248 (1913).
(16) Bragg, W. L. *Proc. roy. Soc.* A, **123**, 537 (1929).
(17) Bragg, W. L. *Z. Kristallogr.* **74**, 237 (1930).
(18) Bragg, W. L., Darwin, C. G. and James, R. W. *Phil. Mag.* (7), **1**, 897 (1926).
(19) Bragg, W. L., James, R. W. and Bosanquet, C. H. *Phil. Mag.* (6), **41**, 309 (1921); **42**, 1 (1921); **44**, 433 (1922).
(20) Bragg, W. L. and Lipson, H. *Z. Kristallogr.* **95**, 323 (1936).
(21) Bragg, W. L. and West, J. *Z. Kristallogr.* **69**, 118 (1929).
(22) Bragg, W. L. and West, J. *Phil. Mag.* (7), **10**, 823 (1930).
(23) Brindley, G. W. *Nature, Lond.*, **137**, 315 (1936).
(24) Compton, A. H. *Phys. Rev.* **9**. 29 (1917).
(25) Compton, A. H. and Allison, S. K. *X-Rays in Theory and Experiment* (1935). London: Macmillan and Co.
(26) Cork, J. M. *Phil. Mag.* (7), **4**, 688 (1927).
(27) Cox, E. G. and Shaw, W. F. B. *Proc. roy. Soc.* A, **127**, 71 (1930).
(28) Darwin, C. G. *Phil. Mag.* (6), **27**, 315, 675 (1914).
(29) Dawton, R. H. V. M. *J. sci. Instrum.* **14**, 198 (1937).
(30) Debye, P. *Ann. Phys., Lpz.*, **43**, 49 (1914).
(31) Debye, P. and Scherrer, P. *Phys. Z.* **17**, 277 (1916).
(32) Dickinson, R. G. and Raymond, A. L. *J. Amer. chem. Soc.* **45**, 22 (1923).
(33) Duane, W. *Proc. nat. Acad. Sci. Wash.* **11**, 489 (1925).
(34) Epstein, P. S. and Ehrenfest, P. *Proc. nat. Acad. Sci., Wash.*, **10**, 133 (1924).
(35) Ewald, P. P. *Ann. Phys., Lpz.*, **49**, 1, 117 (1916); **54**, 519, 557 (1917).
(36) Ewald, P. P. *Handbuch der Physik*, **23**, 208–476 (1933).
(37) Fankuchen, I. *Z. Kristallogr.* **94**, 212 (1936).
(38) Fedorov, E. S. *Bull. Acad. Sci. St-Pétersb.* **10**, 359, 435, 547, 1675 (1916).
(39) Fermi, F. *Z. Phys.* **48**, 73 (1928).
(40) Friedel, G. *C.R. Acad. Sci., Paris*, **157**, 1533 (1913); **169**, 1147 (1919).
(41) Friedrich, W., Knipping, P. and Laue, M. *S. B. bayer. Akad. Wiss.* p. 303 (1912).
(42) Hagg, G. and Soderholm, G. *Z. phys. Chem.* B, **29**, 88 (1935).

(43) HARKER, D. *J. chem. Phys.* **4**, 381 (1936).
(44) HARTREE, D. R. *Phil. Mag.* (6), **50**, 289 (1925).
(45) HARTREE, D. R. *Proc. Camb. phil. Soc.* **24**, 89, 111, 426 (1927); **25**, 310 (1929).
(46) HAVIGHURST, R. J. *Proc. nat. Acad. Sci., Wash.*, **11**, 502 (1925).
(47) HAVIGHURST, R. J. *Phys. Rev.* **29**, 1 (1927).
(48) HELMHOLTZ, L. *J. chem. Phys.* **3**, 740 (1935); **4**, 316 (1936).
(49) HENDRICKS, S. D. *Z. Kristallogr.* **84**, 85 (1932).
(50) HERMANN, C. *Z. Kristallogr.* **68**, 257 (1928).
(51) HILTON, H. *Mathematical Crystallography* (1903). Oxford: Clarendon Press.
(52) HULL, A. W. *Phys. Rev.* **10**, 661 (1917).
(53) *International Tables for the Determination of Crystal Structure* (1935). London: Bell and Sons.
(54) JAMES, R. W. *Z. Kristallogr.* **89**, 295 (1934).
(55) JAMES, R. W. and BRINDLEY, G. W. *Phil. Mag.* (7), **12**, 81, 729 (1931).
(56) JAMES, R. W., WALLER, I. and HARTREE, D. R. *Proc. roy. Soc.* A, **109**, 598 (1925).
(57) KRISHNAN, K. S. and BANERJEE, S. *Philos. Trans.* A, **231**, 235 (1933); **232**, 99 (1933); **234**, 265 (1935); **235**, 343 (1936).
(58) LENNARD-JONES, J. E. *Proc. roy. Soc.* A, **158**, 280, 297 (1937).
(59) LINSTEAD, R. P. *J. chem. Soc.* p. 1016 (1934).
(60) LINSTEAD, R. P. and ROBERTSON, J. M. *J. chem. Soc.* p. 1736 (1936).
(61) LIPSON, H. *Proc. roy. Soc.* A, **151**, 347 (1935).
(62) LIPSON, H. and BEEVERS, C. A. *Proc. roy. Soc.* A, **148**, 664 (1935).
(63) LIPSON, H. and BEEVERS, C. A. *Proc. phys. Soc.* A, **48**, 772 (1936).
(64) LLEWELLYN, F. J., COX, E. G. and GOODWIN, T. H. *J. chem. Soc.* p. 883 (1937).
(65) LONSDALE, K. *Proc. roy. Soc.* A, **123**, 494 (1929).
(66) LONSDALE, K. *Structure Factor Tables* (1936). London: Bell and Sons.
(67) LONSDALE, K. *Proc. roy. Soc.* A, **159**, 149 (1937).
(68) MAUGUIN, C. *Z. Kristallogr.* **76**, 542 (1931).
(69) MULLER, A. *Proc. roy. Soc.* A, **120**, 437 (1928).
(70) NIGGLI. *Geometrische Kristallographie des Diskontinuums* (1919). Leipzig.
(71) NISHIKAWA, S. *Proc. Tokyo math.-phys. Soc.* **8**, 199 (1915).
(72) PATTERSON, A. L. *Phys. Rev.* **46**, 372 (1934).
(73) PATTERSON, A. L. *Z. Kristallogr.* **90**, 517 (1935).
(74) PATTERSON, A. L. *Z. Kristallogr.* **90**, 543 (1935).
(75) PAULING, L. *J. Amer. chem. Soc.* **45**, 2777 (1923).
(76) PAULING, L. *J. Amer. chem. Soc.* **46**, 2738 (1924).
(77) PAULING, L. and SHERMAN, J. *Z. Kristallogr.* **81**, 1 (1932).
(78) PAULING, L. and WILSON, E. B. *Introduction to Quantum Mechanics* (1936). London: McGraw-Hill Co.
(79) PENNEY, G. W. *Proc. roy. Soc.* A, **158**, 306 (1937).
(80) POLANYI, M. *Z. Phys.* **7**, 149 (1921).
(81) POSNJAK, E. and BARTH, T. F. W. *Phys. Rev.* **38**, 2234 (1931).
(82) ROBERTSON, J. M. *Chem. Rev.* **16**, 417 (1935).
(83) ROBERTSON, J. M. *Proc. roy. Soc.* A, **157**, 79 (1936).
(84) ROBERTSON, J. M. *Proc. roy. Soc.* A, **140**, 79 (1933); **142**, 674 (1933).
(85) ROBERTSON, J. M. *Phil. Mag.* (7), **18**, 729 (1934).
(86) ROBERTSON, J. M. *J. chem. Soc.* p. 615 (1935); p. 1195 (1936).
(87) ROBERTSON, J. M. *Nature, Lond.*, **138**, 683 (1936).
(88) ROBERTSON, J. M. *Phil. Mag.* (7), **21**, 176 (1936).
(89) ROBERTSON, J. M. and WOODWARD, I. *J. chem. Soc.* p. 219 (1937).
(90) ROBERTSON, J. M. and WOODWARD, I. *J. chem. Soc.* p. 1817 (1936).
(91) ROBINSON, B. W. *Proc. roy. Soc.* A, **142**, 422 (1933).
(92) ROBINSON, B. W. *J. Sci. Instrum.* **10**, 233 (1933).
(93) SCHIEBOLD, E. *Die Lauemethod*, Akad. Verlagsgesellschaft, Leipzig (1932).
(94) SCHOENFLIES, A. *Z. Kristallogr.* **54**, 545 (1915).
(95) SHEARER, G. *Proc. phys. Soc.* **35**, 81 (1922).

(96) SHEARER, G. *Proc. roy. Soc.* A, **108**, 655 (1925).
(97) "Strukturbericht." *Z. Kristallogr. Erganzungsband*, **1–3** (1913–35); also sections of *Z. Kristallogr. Referantenteil*, **1–4**, still continuing.
(98) THOMAS, L. H. *Proc. Camb. phil. Soc.* **23**, 542 (1926).
(99) WALLER, I. *Z. Phys.* **17**, 398 (1923).
(100) WALLER, I. *Ann. Phys., Lpz.*, **83**, 153 (1927).
(101) WARREN, B. E. and GINGRICH, N. S. *Phys. Rev.* **46**, 368 (1934).
(102) WEISSENBERG, K. *Z. Phys.* **23**, 229 (1924).
(103) WEST, J. *Z. Kristallogr.* **74**, 306 (1930).
(104) WOOSTER, W. A. *Z. Kristallogr. Referatenteil*, **3**, 105 (1931).
(105) WOOSTER, W. A. and ROBINSON, B. W. *Nature, Lond.*, **131**, 545 (1933).
(106) WYCKOFF, R. W. G. *The Analytical Expression of the Results of the Theory of Space Groups* (1930). Carnegie Institute of Washington.
(107) WYCKOFF, R. W. G. *The Structure of Crystals* (1931; Supplement, 1935). New York: Chemical Catalog Co.
(108) ZACHARIASEN, W. H. *Z. Kristallogr.* **89**, 442 (1934).

DIAMAGNETIC AND PARAMAGNETIC ANISOTROPY OF CRYSTALS

By KATHLEEN LONSDALE, D.Sc.
Davy Faraday Laboratory, Royal Institution, London, W 1

§ 1. INTRODUCTION

Since the early experiments of Plücker, Faraday[12] and Tyndall[64] it has been well known that non-cubic single crystals, even if cut to a spherical shape, tend to assume a preferred orientation in a magnetic field. The strength of the effect, all other factors being constant, is characteristic of the crystalline substance used and of its initial orientation. Such crystals are said to exhibit *magnetic anisotropy*. Faraday defined a *magnecrystallic substance* as "one which in the crystalline state could conduct onwards or permit the exertion of the magnetic force with more facility in one direction than another". Such a substance, when placed in a magnetic field, will experience a force tending to turn it into the position of minimum potential energy.

Faraday himself hoped to find some relation between magnetism and the forces which build up crystal structure, but the experimental means at his disposal and the knowledge of crystal structure available at that time were far too meagre for his hopes to come to fruition. Even eighty years later, when X-ray methods of investigating the structure of crystals had been in use for some fifteen years, little had been done to explain the behaviour of crystals in a magnetic field in terms of their structure[5]. This was partly because, although in the crystallographic description of any substance the optical constants would be recorded almost as a matter of course, the analogous data for the magnetic constants were almost entirely lacking.

During the last few years, however, there has been a revival of interest in this branch of physics, and great advances have been made in both theoretical and practical knowledge.

In the following article the various factors affecting the behaviour of a crystalline body placed in a magnetic field will first be considered, and then the methods by means of which the magnetic constants may be measured. Finally, the results of actual measurements will be discussed, together with their theoretical explanation and their bearing on modern problems of the electronic structure of matter.

§ 2. FACTORS AFFECTING THE BEHAVIOUR OF A CRYSTALLINE SUBSTANCE WHEN PLACED IN A MAGNETIC FIELD

The behaviour of a substance placed in a magnetic field depends on (*a*) the nature of the substance, whether diamagnetic or paramagnetic relative to the surrounding medium, (*b*) the shape of the specimen, (*c*) the uniformity of the magnetic field, (*d*) the anisotropy of the substance, (*e*) the temperature, and (*f*) the field strength.

These factors assume varying importance for different substances and cannot, in general, be separately considered as they are inter-related the one with the other. They should all, however, be recorded in any description of magnecrystallic behaviour.

Diamagnetism. Any substance placed in a magnetic field experiences a feeble induced opposing magnetization I (I = magnetic moment/unit volume) which is proportional to the field H. The factor κ in the equation

$$I = \kappa H$$

is the *diamagnetic susceptibility per unit volume* of the substance, and it is intrinsically negative. Its magnitude, which is usually of the order of 10^{-6} c.g.s.e.m.u., is dependent upon the distribution of electron density in the substance and is independent of temperature, field strength and other external factors, except in so far as these may affect the electronic state of the material. The classical Langevin-Larmor theory states that the effect of an imposed magnetic field is to accelerate or retard the movement of the electrons in their orbits, the electronic currents thus induced producing a field which is opposite in direction to the inducing field. This theorem gives for the *gram-atomic susceptibility* χ_A (where $\chi_A = \kappa . \times$ atomic weight/density) of an atom having spherical symmetry,

$$\chi_A = -\frac{Ne^2}{6mc^2} \sum_n \overline{r^2} = -2{\cdot}83 \times 10^{10} \sum_n \overline{r^2} \qquad \text{.....(i),}$$

where the summation is taken over all the mean squares of the orbital electronic radii.

This result is not modified by quantum mechanics except that it now refers to the mean susceptibility for an aggregate of atoms having statistical spherical symmetry, and simply gives an estimate of the average electron density distribution.

Paramagnetism. If the substance is composed of ions, atoms or molecules having a resultant orbital or spin moment, the tendency of an external magnetic field is to align these *molecular magnets* is such a way that their potential energy in the field is a minimum. Complete alignment is opposed by the kinetic energy of the molecules themselves, by interaction with neighbouring molecules or ions, and by *freezing* or *quenching* of the moments owing to the presence of a local electric field, such as occurs in crystals. This paramagnetic effect, when it exists, is generally much greater than the universal diamagnetic effect and more or less completely masks it. For paramagnetic substances the factor $\kappa\,(=\kappa_P - \kappa_D)$ in the equation $I = \kappa H$ is therefore positive, since the resultant induced magnetization is in the same sense as the external field. Except for low temperatures and strong fields, the *paramagnetic susceptibility* κ_P is independent of field strength, but it is not independent of temperature. Classical theory leads to the Curie-Weiss expression

$$\chi_M = \frac{C_M}{T - \theta} \qquad \text{.....(ii),}$$

where χ_M is the gram-molecular susceptibility, C_M the Curie constant, T the absolute temperature and θ the Curie temperature or Curie point. Below θ the thermal

motion is not vigorous enough to overcome the tendency to alignment of the molecular magnets, hysteresis appears and the substance is *ferromagnetic*. Ferromagnetism, whether of isotropic or of anisotropic substances, is, however, outside the scope of the present article, and in all that follows it will be assumed that the substance is at a temperature higher than this critical temperature θ.

Experimentally it is found that $C_M = 0{\cdot}1241\ p_B^2$, where p_B is the magnetic moment of the molecule expressed in terms of *Bohr magnetons* (unit $he/4\pi mc$).

Influence of electronic state. Ions or molecules showing only a diamagnetic effect are characterized by the presence of closed systems (that is, states for which the resultant moment $j=0$). The wide prevalence of diamagnetism in salts and in complex compounds is due to the common occurrence of ions or groupings having configurations similar to those of the inert gases, of type $^1S_0$, having no resultant spin or orbital moment. The majority of molecules, and especially of organic molecules, contain an even number of electrons, whose spins are balanced in pairs, while the orbital moments either compensate each other or are in some other way ineffective in responding to an applied field. Aggregates of such molecules will be diamagnetic, although for molecules in which there is a large departure from central symmetry a small modifying paramagnetic effect may come into action.

Paramagnetism is found chiefly among salts containing ions of the transition or rare earth elements, having incomplete electron groups, and it is also found in the comparatively few molecules containing an odd number of electrons. Simple quantum considerations[58] show that when an atom has zero orbital angular momentum (that is, is in an S state) its magnetic moment p_B is given by

$$p_B = \sqrt{\{4s\,(s+1)\}} \qquad \text{.....(iii),}$$

thus depending only on the value s of the spin moment. This is known as the *spin-only value*.

More generally, when both orbital (l) and spin (s) moments are present and are coupled, they give rise to a resultant mechanical moment j, and a magnetic moment[23]

$$p_B = g\ \sqrt{\{j\,(j+1)\}} \qquad \text{.....(iv),}$$

where g is the Landé splitting factor. This formula, however, only applies when the energy separation of the levels near the ground state is large compared with kT ($h\Delta\nu \gg kT$). If the multiplet intervals are small ($h\Delta\nu \ll kT$) there may be a distribution of magneton values over a certain range, all the ions not being in the lowest state. The spin and orbital moments may then be quantized separately, leading to the formula

$$p_B = \sqrt{\{4s\,(s+1) + l\,(l+1)\}} \qquad \text{.....(v).}$$

Empirically it is found that when $h\Delta\nu \ll kT$, the p_B values lie between the limits[60]

$$\sqrt{\{4s\,(s+1)\}} \quad \text{and} \quad \sqrt{\{4s\,(s+1) + l\,(l+1)\}}.$$

This is due to interaction effects[61] which render the orbital l moment partially or totally ineffective in responding to an external field, though the spin s moment is comparatively free to do so. Such effects, which will be considered in more detail

later, are particularly important in dealing with paramagnetic crystals. They also lead to some curious complications, as an atom in the free state will sometimes behave very differently from the same atom in bulk, while different allotropic forms of the same substance may have completely different mean susceptibilities[3, 8, 52].

Changes of susceptibility on melting. A certain amount of conflicting evidence exists as to the changes in diamagnetism that take place at the melting-point. In 1914 Oxley[43] reported that, in general, a decrease in diamagnetic susceptibility took place on crystallization of organic compounds, presumably due to a decrease in the sum of the projected areas of the electron orbits. The change was very small for aliphatic compounds, but of the order of 5 per cent for aromatics. Doubt, however, was cast on the soundness of these results by Krishnan, Guha and Banerjee[31] who showed that if the susceptibility in the solid state is taken as the mean of the three principal susceptibilities of the single crystals, the change on melting, $\Delta\chi/\chi$, becomes much smaller (0·5 per cent for naphthalene instead of Oxley's 5 per cent). It therefore appears that Oxley's solids may have been subject to some preferential orientation of the microcrystals.

Cabrera and Fahlenbrach[7] found a change of diamagnetism on solidification which bore a simple exponential relationship to the electric dipole moment of the substance, but Bose[4], as the result of measurements on twenty-seven aromatic compounds, came to the conclusion that, within the limits of experimental error, the diamagnetic susceptibility of an aromatic compound is the same, whether it is single crystal, liquid or in solution.

In some cases, however, there is a definite and striking change of diamagnetism at the melting-point, and for such substances it is clear that the susceptibility in the solid state must be intimately connected with the crystal structure and not merely a property of the individual atoms. The mean mass susceptibility χ_m ($\chi_m = \kappa/\text{density}$) of bismuth just below the melting-point is $-1{\cdot}02 \times 10^{-6}$, while that of molten Bi is only $-0{\cdot}08 \times 10^{-6}$. Similar but less striking changes take place for other metals and semi-metals[62]. Tin is of particular interest in that the susceptibility of white tin (metallic binding, tetragonal structure) is positive, $\chi_m = +0{\cdot}036 \times 10^{-6}$, while that of grey tin (homopolar binding, tetrahedral cubic structure) is negative, $\chi_m = -0{\cdot}35 \times 10^{-6}$. On melting, white tin becomes feebly diamagnetic, $\chi_m = -0{\cdot}046 \times 10^{-6}$. According to Rao[52] the paramagnetism of white tin in bulk diminishes with decreasing particle size, slowly at first, but more rapidly when the particle diameter becomes less than $2{\cdot}0\mu$. At about $1{\cdot}4\mu$ the susceptibility is zero, and at lower diameters it becomes negative, reaching a value of $-0{\cdot}18 \times 10^{-6}$ for diameter $0{\cdot}4\mu$. The extrapolated value for χ_m for particles of *atomic* diameter is about $-0{\cdot}30 \times 10^{-6}$, nearly the same value as that found for grey tin, or calculated for the atom from its electron distribution. On melting and recrystallizing the finest particles the paramagnetic value is regained, showing it to be a property of the crystalline state and not of the individual atoms.

It is clear that these facts present problems the solution of which requires a thorough knowledge both of the magnetic properties and of the electronic structure of the solid state.

Influence of shape of specimen and uniformity of field. A sphere of isotropic material will not be acted upon by any magnetic force in a strictly homogeneous field. If the field is not uniform the sphere will experience a lateral force tending to move it to the weakest part of the field if the substance is diamagnetic relative to the surrounding medium, or to the strongest part of the field if the substance is paramagnetic relative to the surrounding medium.

The behaviour of an isotropic body of asymmetric shape is rather more complicated. In a field which was strictly uniform prior to the introduction of the body there is an exceedingly weak force tending to set a rod of material, whether diamagnetic or paramagnetic, with its greatest length along the lines of force. This effect is, however, so minute and so difficult to separate from others that it has almost certainly never been directly observed[5, 29]. If the field is not uniform, an asymmetric body will not only experience the lateral force mentioned above, according to its dia- or paramagnetism, but also a torque, tending to bring its direction of greatest length into such a position that the demagnetizing force is a minimum. An isotropic rod of diamagnetic material will try to set itself equatorially, and an isotropic rod of paramagnetic material will try to set itself axially with respect to the lines of force. This tendency may be used to test the homogeneity of a magnetic field[31].

All the above effects will be superimposed upon and may mask the effects due to crystalline structure which will now be considered.

Influence of crystal structure. When a crystalline body is placed at random in a magnetic field, the direction of magnetization does not, in general, correspond with the direction of the applied field. There are, however, three mutually perpendicular principal directions along which the two are coincident; these are the *axes of principal magnetization.* Just as in the case of optical and elastic properties, the susceptibilities of a single crystal in different directions can be represented by radii vectors drawn from a common centre, the end-points of which define an ellipsoid—the *ellipsoid of magnetic induction.* In the case of a cubic crystal this ellipsoid degenerates into a sphere and the substance is isotropic[69]. A substance may also be isotropic because it is composed of microcrystals of absolutely random orientation.

The axes of the ellipsoid of magnetic induction are the axes of principal magnetization, and the susceptibilities along these directions are the *principal susceptibilities* of the crystal (volume susceptibilities $\kappa_1, \kappa_2, \kappa_3$; mass susceptibilities χ_1, χ_2, χ_3. The gram-molecular susceptibilities are also frequently denoted by χ_1, χ_2, χ_3). For hexagonal, tetragonal and trigonal crystals the magnetic ellipsoid is an ellipsoid of rotation, one of the two principal susceptibilities ($\chi_{\|}$) lying along the principal axis of the crystal. In the orthorhombic system the axes of the magnetic ellipsoid, which are all unequal, lie along the crystalline axes; in the monoclinic system one principal axis is coincident with the symmetry axis of the crystal (the *b* axis), while the other two lie in the (010) plane; in the triclinic system the directions of principal susceptibility are not directly related to the crystallographic axes.

The absolute values of the principal susceptibilities are very simply related to the mean susceptibility as determined from a crystal powder:

$$3\bar{\chi} = \chi_1 + \chi_2 + \chi_3 \qquad \text{.....(vi).}$$

From table 1 it will be seen that different substances show widely different degrees of anisotropy even when they are isomorphous in structure. The anisotropy $\Delta\chi$ is here defined as the difference in the values of two principal susceptibilities, although sometimes in the literature it is expressed as the ratio of two susceptibilities.

Table 1

Substance and crystallography	Mean gram-mol. susceptibility $\bar{\chi}$ ($\times 10^6$)	Gram-mol. anisotropy at room temperature $\Delta\chi$ ($\times 10^6$)
Quartz, SiO_2 (trigonal)	− 27·6	$\chi_1 - \chi_\perp =$ − 0·12
$BaCO_3$ (orthorhombic)	− 58·8	$\chi_a - \chi_c =$ 4·93 $\chi_b - \chi_c =$ 5·07
$BaSO_4$ (orthorhombic)	− 73·9	$\chi_a - \chi_c =$ − 0·62 $\chi_b - \chi_c =$ 0·72
Urea, CH_4N_2O (tetragonal)	− 33·6	$\chi_1 - \chi_\perp =$ 1·05
Oxalic acid, $(COOH)_2\,2H_2O$ (monoclinic)	− 56·1	$\chi_1 - \chi_2 =$ 1·91 $\chi_1 - \chi_3 =$ 7·05
Anthracene, $C_{14}H_{10}$ (monoclinic)	− 130·1	$\chi_1 - \chi_2 =$ 136·3 $\chi_1 - \chi_3 =$ 27·4
Metal-free phthalocyanine, $C_{32}N_8H_{18}$ (monoclinic)	− 422	$\chi_1 - \chi_2 =$ 372 $\chi_1 - \chi_3 =$ 397
Graphite, C (hexagonal)	− 92	$\chi_3 - \chi_\perp =$ − 258
Bismuth, Bi (trigonal)	− 280	$\chi_3 - \chi_\perp =$ 92
Zinc, Zn (hexagonal)	− 10·3	$\chi_3 - \chi_\perp =$ − 2·9
White tin, Sn (tetragonal)	+ 3·09	$\chi_3 - \chi_\perp =$ − 0·34
$FeCO_3$ (trigonal)	+42690	$\chi_1 - \chi_\perp =$ −6870

Tutton series of isomorphous monoclinic salts (selected examples):

Substance	Mean gram-mol. susceptibility $\bar{\chi}$ ($\times 10^6$)	Gram-mol. anisotropy at room temperature $\Delta\chi$ ($\times 10^6$)
$FeSO_4(NH_4)_2SO_4.6H_2O$	+11480	$\chi_1 - \chi_2 =$ 2582 $\chi_1 - \chi_3 =$ 213
$CoSO_4(NH_4)_2SO_4.6H_2O$	+ 9590	$\chi_1 - \chi_2 =$ 3023 $\chi_1 - \chi_3 =$ 1541
$NiSO_4(NH_4)_2SO_4.6H_2O$	+ 3967	$\chi_1 - \chi_2 =$ 110 $\chi_1 - \chi_3 =$ 106
$MnSO_4(NH_4)_2SO_4.6H_2O$	+13830	$\chi_1 - \chi_2 =$ 8·7 $\chi_1 - \chi_3 =$ 6·0
$MnSeO_4.Rb_2SeO_4.6H_2O$	+13980	$\chi_1 - \chi_2 =$ 9·0 $\chi_1 - \chi_3 =$ 7·3
$ZnSO_4(NH_4)_2SO_4.6H_2O$	Diamagnetic	$\chi_1 - \chi_2 =$ 1·2 $\chi_1 - \chi_3 =$ 0·8

§ 3. METHODS OF DETERMINING CRYSTAL SUSCEPTIBILITIES

Measurements of crystal susceptibilities involve (*a*) the location of the principal axes, (*b*) the determination of χ along those axes, or (*c*) the determination of χ along one axis, or of $\bar{\chi}$, together with the measurement of the anisotropy $\Delta\chi$ in two planes.

Location of the principal axes. In the case of crystals belonging to the orthogonal systems, the magnetic axes coincide with the crystallographic axes. If the crystals have well developed faces, the symmetry axes may usually be located by an external examination; sometimes they may be found optically, by observation of the extinction directions. Single metal crystals are often grown in the form of long rods; methods of dealing with these have been adequately described in a previous *Report*[1]. Monoclinic crystals have only one symmetry axis (b); the principal susceptibility along this direction is usually denoted by χ_3; χ_1 is the algebraically greater, χ_2 the lesser of the two principal susceptibilities in the (010) plane. The position of χ_1 is further defined by saying that χ_1 makes an angle ψ with the c axis, $\beta-\psi$ with the a axis, where β is the obtuse angle between c and a. Monoclinic crystals therefore have four magnetic constants, χ_1, χ_2, χ_3 and ψ. Measurements of absolute susceptibility along χ_3 and along three other known directions in the crystal will give all four constants, but this indirect way of determining ψ is not satisfactory in practice[29]. A much better method is that of suspending the crystal in a strictly uniform horizontal field with its b axis vertical, and rotating it until its suspension is free from torsion. The χ_1 axis will then lie along the direction of the field, and its inclination to any natural vertical face (hol) of the crystal can be measured directly. A neat and very accurate method of doing this has been described by Krishnan and Banerjee[27, 29]. The directions of the magnetic axes in triclinic crystals can be found in the same direct way, although the procedure is somewhat more complicated, since none of the axes need coincide with any crystallographic axis[32].

Determination of absolute susceptibility. The methods generally used for absolute measurements of susceptibility are based on that of Faraday, who determined the force F_x on a specimen placed in a non-uniform field:

$$F_x=\chi_m m H_y \frac{dH_y}{dx} \qquad \text{.....(vii).}$$

By using a comparison method and arranging that the displacements shall be small, errors due to the difficulty of measuring $H\dfrac{dH}{dx}$ and its variations from point to point are eliminated. Sucksmith[63] introduced a neat optical device for magnifying the displacements, which could then be observed down to 10^{-5} cm., and this method was developed by Jackson[24] and applied by him to low-temperature work, mainly on paramagnetic crystals, and by Shoenberg and Uddin[55] to work on Bi and Bi alloys.

Rabi[50] has described a useful null method of measuring absolute susceptibilities, which forms the basis of much of the most accurate work on small single crystals that is now being carried on. It depends on the fact that when the volume susceptibility of the medium surrounding a crystal is equal to the volume susceptibility of the crystal in any particular direction, then there will be no lateral force acting on the crystal in that direction, even in a non-uniform field. Rabi suspended his crystal vertically in a non-homogeneous field by a glass thread about

100 cm. long and just sufficiently thick to prevent any detectable turning of the crystal due to its shape or anisotropy. He surrounded it by a liquid bath whose susceptibility could be varied at will, within the range of susceptibility of the crystal. By means of a torsion head the crystal could be turned so as to have a particular direction along the field direction X. In general there will be a lateral motion of the crystal in the Y direction (as far as the bending moment of the glass thread allows) when the field is put on. The concentration of the surrounding solution is varied until there is no motion of the crystal in the Y direction; the volume susceptibility of a sample of the solution may be determined by any standard method, and this then gives the volume susceptibility of the crystal in the X direction.

Krishnan and his collaborators[31] used a very fine quartz fibre in place of the glass thread, giving much greater accuracy for the measurement of susceptibility

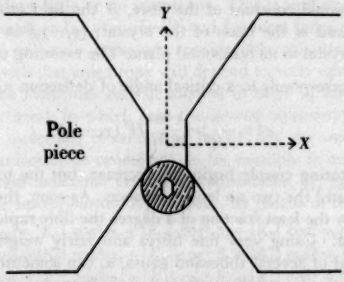


Figure 1. Horizontal plane.

in that direction which naturally sets itself along the field. They later[26] dispensed with the fibre altogether by the very neat device of suspending the crystal in a liquid bath of its own density, but of variable susceptibility. The latter method requires careful and patient manipulation and is not suitable for substances whose density or susceptibility is outside the usual range of aqueous solutions, but it gives reliable results even for very minute crystals, on a wide range of materials.

Measurement of anisotropy. The measurement of crystal anisotropy by the determination of absolute susceptibilities alone is not very accurate. It is much more satisfactory to use methods in which the force exerted on the crystal is proportional to the *difference* of susceptibilities in two orthogonal directions.

One such method, which was used by Krishnan and his collaborators in much of their earlier work[29, 31], has been already described in these *Reports*[1]. It utilizes the change in the period of oscillation of a single crystal suspended by a vertical torsion fibre, when placed in a uniform horizontal magnetic field.

A second method, also developed by the same workers[27], is now extensively used for measurements on small single crystals. As in the first method, complications due to lateral forces acting on the crystal or measurable torque due to shape are avoided by using a field as nearly uniform as possible. The crystal, in its natural form, is suitably suspended from a fine quartz fibre attached to a torsion head, which is then rotated until the axis of greatest magnetic susceptibility χ_1 (say) lies along the field and the fibre is torsionless. If the torsion head be slowly rotated from the equilibrium setting by an angle α, the crystal will also rotate in the same direction but by a much smaller angle ϕ, the twist on the fibre being balanced by a restoring couple due to the field and the crystal anisotropy. The relation governing the motion is

$$c(\alpha-\phi)=\frac{1}{2}\frac{m}{M}H^2(\chi_1-\chi_2)\sin 2\phi \qquad \text{.....(viii),}$$

where c is the torsional constant of the fibre, H the field strength, M the gram-molecular weight and m the mass of the crystal; $(\chi_1-\chi_2)$ is the gram-molecular anisotropy of the crystal in its horizontal plane. The restoring couple is a maximum when $\phi=\frac{\pi}{4}$; this corresponds to a critical angle of deflection α_c, such that

$$c\left(\alpha_c-\frac{\pi}{4}\right)=\frac{1}{2}\frac{m}{M}H^2(\chi_1-\chi_2) \qquad \text{.....(ix).}$$

Thereafter the restoring couple begins to decrease, but the torsional couple continues to increase, and the two no longer balance. As soon, therefore, as the value α_c is overreached by the least fraction of a degree the fibre rapidly untwists and the crystal spins round. Using very fine fibres and fairly weighty crystals of large anisotropy in a field of several thousand gauss, α_c can sometimes amount to 20 or 30 complete revolutions, and the consequent spinning round of the crystal when $\phi=\frac{\pi}{4}$ is most dramatic.

If the mean susceptibility is large and the anisotropy small (as for example in the case of manganous salts; see table 1) the effect of asymmetry of shape and non-homogeneity of field may be important, even in the most uniform field obtainable[29]. In order to eliminate these effects, the crystals may be surrounded by a liquid bath of their own mean volume susceptibility.

Really accurate values of the magnetic constants are as yet available for relatively few crystals, even among those of known structure. Measurements of mean susceptibility $\bar{\chi}$ have, however, been made on the powdered crystals of a great many substances.

§ 4. ANISOTROPY OF DIAMAGNETIC CRYSTALS

Diamagnetism of molecular and ionic compounds is inherently a property of the electron distribution in individual molecules or ions and is, as Pascal pointed out some thirty years ago, an *additive property* of organic compounds[44]. The

molecular susceptibility χ_M of a compound having the formula $A_\alpha B_\beta C_\gamma \ldots$ is given by the equation

$$\chi_M = \Sigma\alpha . \chi_A + \lambda \qquad \ldots\ldots(x),$$

where α, β, ..., are the number of atoms of susceptibilities χ_A, χ_B, ... and λ is a constitutive correction factor, dependent on the nature of the chemical linkages. When it is remembered that χ_M is the mean of the three principal susceptibilities in the case of solid compounds, it is clear that the mutual influences of neighbouring molecules in the solid state must be negligible. In other words, the molecules or ions in the solid state are themselves magnetically anisotropic and the observed principal susceptibilities of the crystal are simply the result of adding (according to the rules for the combination of tensors) the susceptibilities of the individual molecules which compose it[40]. If the electron distribution in the molecule is such that the molecule is approximately isotropic from a magnetic point of view, then the crystal will also show little or no anisotropy. This is said to be the case for the few aliphatic compounds so far examined[2] and for most inorganic compounds.

If, however, the molecule possesses considerable anisotropy, the degree to which the crystal reveals that anisotropy will depend entirely upon the arrangement of the molecules. The molecules may be aligned in more or less parallel positions (layer or chain structures), in which case the crystal anisotropy will be nearly as large as that of the molecule; it cannot be larger. Or the molecules may be so arranged that their anisotropies cancel out, as for example in cubic crystals.

The principal gram-molecular crystal susceptibilities, χ_1, χ_2, χ_3, are related to the *principal molecule susceptibilities*, K_1, K_2, K_3, by means of the three equations

$$\left.\begin{aligned} \chi_1 &= |K_1| \cos^2\alpha_1 + |K_2| \cos^2\beta_1 + |K_3| \cos^2\gamma_1 \\ \chi_2 &= |K_1| \cos^2\alpha_2 + |K_2| \cos^2\beta_2 + |K_3| \cos^2\gamma_2 \\ \chi_3 &= |K_1| \cos^2\alpha_3 + |K_2| \cos^2\beta_3 + |K_3| \cos^2\gamma_3 \end{aligned}\right\} \qquad \ldots\ldots(xi),$$

where $\cos\alpha_1$, $\cos\beta_1$, $\cos\gamma_1$ are the direction cosines of K_1, K_2, K_3 with respect to χ_1, etc. The magnitude and direction of χ_1, χ_2, χ_3 may be determined experimentally as in § 3, but the positions of the principal molecule susceptibilities are not known except for triclinic crystals, where they must necessarily coincide with χ_1, χ_2, χ_3. An X-ray analysis will, however, give the direction cosines of certain important orthogonal directions in the molecule and equations (xi) may then be used to give the values of the molecule susceptibilities along those directions, if the axes of reference are first transformed (if necessary) from the crystallographic to the principal magnetic axes of the crystal. In uniaxial crystals there are only two principal susceptibilities, $\chi_\parallel$ and $\chi_\perp$, and these are insufficient for the calculation of the three molecule susceptibilities unless it can safely be assumed that two of the latter are equal.

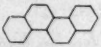
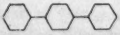
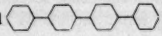
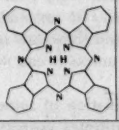
Inorganic and organic salts. Examination of some of the commoner inorganic salts has shown that carbonates and nitrates possess a relatively high anisotropy as compared with the corresponding sulphates (see table 1)[67,31], and that the numerically greatest diamagnetic susceptibility is perpendicular to the plane of the nitrate and carbonate radicles. Studies of some diamagnetic double salts[28] have

also shown that whereas sulphates and selenates have very feeble anisotropies, those of the oxalates, though still small, are considerably larger.

Oxalic acid dihydrate[37] also has a relatively large anisotropy and in this case again the greatest diamagnetic susceptibility is that normal to the plane of the oxalate group as revealed by X-ray analysis[53]. The sulphate and selenate groups are not plane, but have approximately tetrahedral symmetry.

Ring compounds. The aromatic organic compounds show really striking anisotropy[2,31,27,40] as do also compounds containing cyanuric or phthalocyanine rings[38]. This shows that the molecules themselves must be remarkably anisotropic; data deduced for some individual molecules from magnetic measurements on crystals and X-ray analyses of structure are given in table 2[38,40]. X-ray analysis has shown that all these molecules possess a plane configuration; K_1, K_2 are taken in the plane of the molecule, K_3 normal to that plane.

Table 2

Substance and formula	$-K_1 \times 10^6$	$-K_2 \times 10^6$	$-K_3 \times 10^6$
Naphthalene	56·1	53·9	169·0
Anthracene	75·8	62·6	251·8
Hexamethylbenzene	101·1	102·7	163·8
Durene	82·4	77·3	143·9
Chrysene 	88·0	83·3	310·8
p-Diphenylbenzene 	96·8	88·1	271·3
p-Diphenylbiphenyl 	122	110	372
Cyanuric trichloride $C_3N_3Cl_3$	70·9	71·2	101·3
Metal-free phthalocyanine $C_{32}N_8H_{18}$ 	165	120	982

It is clear that in the case of such molecules, where the value of the diamagnetic susceptibility normal to the plane of the molecule is numerically so much greater than either of the other two, there cannot be even an approximation to a spherical

symmetry of electron distribution in the individual atoms. Unless there is such an approximation, however, Langevin's formula (i) does not apply.

Ehrenfest[10] and Raman[51] have suggested that large diamagnetic anisotropy may be associated with the existence of electrons moving in plane orbits of large area; and more recently Pauling[45] has proposed a theory in which he assumes that the *aromatic* or *resonance* electrons (one to each carbon atom) are free to move under the influence of the impressed fields from carbon atom to adjacent carbon atom. The problem of calculating the magnetic anisotropy of such a system is equivalent to that of calculating the magnetic effect of the currents induced in a conducting network. By making certain rather large corrections designed to allow for variations of the electronic density in different compounds and for the effect of curvature of the transit from atom to atom, Pauling arrives at values of the molecular anisotropy for a number of benzene derivatives which are within 10 per cent of the experimental values.

In order to make quantitative use of the experimental results it was independently suggested[38] that an aromatic molecule could be treated as a system of spherically symmetrical atoms, which would account for the whole of the susceptibility *in* the molecular plane, and to which formula (i) might be applied without correction, together with a system of electrons constrained to precess only in plane orbits. The magnetic anisotropy ($\Delta K = K_3 - \frac{1}{2}(K_1 + K_2)$) of the molecule would be due entirely to the latter and the mean area $\bar{S}$ of these plane orbits could be calculated from the formula

$$\Delta K = -\frac{Ne^2}{4\pi mc^2}\sum_p \bar{S} \qquad \text{.....(xii),}$$

where p is the number of resonance electrons in the molecule. On applying these formulae to the compounds in table 2 among others, it was found that the normal valency electrons occupied orbits of effective radius about 0·70 to 0·78 A., agreeing well with the values found from X-ray analysis, while the resonance electrons had an effective orbital radius of 1·50 to 1·60 A., comparable with that of the benzene ring itself. In metal-free phthalocyanine the anisotropy due to the 14 resonance electrons in the central closed ring was such as to indicate that they occupy *molecular orbitals* of area a little larger than that of the central ring itself.

Quantum theory. Even more recently London[36] has provided a quantum mechanical basis for the molecular orbital theory in the case where a magnetic field is present. He uses the method of approximation as applied in Bloch's theory of metals, which corresponds to the method of molecular orbitals in the theory of chemical combination, and shows that when two conditions are fulfilled, namely (1) the existence of a closed chain of equivalent pairs of atoms, (2) the association of an uneven number of electrons with each valency bond, an additional susceptibility results, which corresponds to an induced electronic current circulating from one atom to another around the chain. These conditions are fulfilled for aromatic ring-systems and also for cyanuric and phthalocyanine rings. Such induced currents cannot appear in the interior of saturated molecules nor, as intermolecular currents, between saturated molecules. London has succeeded in calculating the

relative anisotropies $\Delta\chi$ of the diamagnetic susceptibilities for various aromatic molecules, taking the anisotropy of benzene ($\Delta\chi_{benz.} = 54 \times 10^{-6}$) as unit. The results* are as shown in table 3.

Table 3

Substance	$\Delta\chi/\Delta\chi_{benz.}$ experimental	$\Delta\chi/\Delta\chi_{benz.}$ theoretical
Naphthalene	2·11	2·18
Anthracene	3·38	3·45
Phenanthrene	3·07	3·19
Pyrene	4·13	4·46
Diphenyl	2·20	2·21

The calculation makes no use of any adjustable parameter and introduces no corrections; it is successful (as Pauling's theory was not) in accounting for the difference between isomers such as anthracene and phenanthrene, and also in showing why the anisotropy of diphenyl should be more than twice that of benzene.

Determination of molecular orientations. The fact that in aromatic compounds (see table 2)

$$|K_3| \gg |K_1| \text{ and } |K_2|$$

and

$$|K_1| \fallingdotseq |K_2|$$

has made it possible for diamagnetic measurements to be used for the determination of molecular orientations in crystals of unknown structure[31, 27, 39, 40]. If $|K_1| = |K_2|$ equations (xi) reduce to

$$|\chi_3| = |K_1| + \Delta K \cos^2 \gamma_1, \text{ etc.},$$

χ_1, χ_2, χ_3 are measurable, $|K_1|$ and ΔK can frequently be estimated with a fair degree of accuracy by the application of Pascal's additive law and the knowledge of ΔK for related aromatic compounds. The angles γ_1, γ_2, γ_3, which give the direction of the *normal* to the molecular plane relative to the principal susceptibility axes of the crystal, can therefore be calculated. This method of magnetic structure analysis has proved a valuable adjunct to X-ray methods, and is capable of extension to other than aromatic compounds. For example, the changes of crystal structure of NH_4NO_3 between 27 and 130° C., have been checked by measurements of magnetic anisotropy over that range of temperature[42].

Although the theories advanced to explain the anisotropy of aromatic molecules have not been quantitatively extended to the case of carbonate, nitrate and oxalate groups, it is clear that the anisotropy of these radicles is also closely connected with the fact that they are plane, conjugated systems.

Graphite, bismuth and some alloys. The abnormal diamagnetism of graphite and of bismuth has long excited interest, especially in view of its temperature variation[30, 35], dependence on particle size[13, 21] and on its chemical treatment, presence of dissolved atoms and so on[13, 53]. The remarkable maximum to be observed in the diamagnetism of the γ-alloys[11] is almost certainly a structural effect,

* Minor corrections have been made in this table in accordance with some private correspondence with Dr London.

although these alloys are cubic and therefore show no anisotropy. Graphite has gram-atomic susceptibilities at room temperature of

$$\chi_{\parallel} = -264 \times 10^{-6}, \quad \chi_{\perp} = -6 \times 10^{-6}$$

(mass susceptibilities -22×10^{-6} and $-0{\cdot}5 \times 10^{-6}$).

The susceptibility $\chi_{\perp}$ in the network of trivalent atoms which form the basal plane is almost exactly the same as that of diamond, $-5{\cdot}8 \times 10^{-6}$, and is relatively unaffected by particle size, chemical treatment, etc. The successive layers of carbon atoms in graphite are separated by the van der Waals' distance 3·41 A., and along this direction the susceptibility is not only exceptionally large but is very dependent on particle size, temperature, etc. The usual explanation for the large value of $\chi_{\parallel}$ is that there are loosely bound electrons executing large orbits in the basal plane, corresponding to the resonance electrons in aromatic compounds. These electrons are also responsible for the semi-metallic properties of graphite.

An explanation of the abnormally high diamagnetism of γ alloys and of some of the semi-metals, notably bismuth, has recently been given in terms of the Bloch theory of metals[47, 25, 68]. The diamagnetic susceptibility of metals is due partly to the core electrons (calculated simply in accordance with equation (i)), and partly to the valency electrons. If the valency electrons are considered to be free, their volume susceptibility can be quite easily calculated by quantum theory. It is found that they possess a resultant *paramagnetism* whose value is given by

$$\kappa = \tfrac{2}{3}\,[2{\cdot}2 \times 10^{-14}\, n^{\frac{1}{3}}] \qquad \text{.....(xiii)},$$

where n is the number of electrons per unit volume. The electrons in a crystal are however not free and their susceptibility may differ considerably from the above value.

In Bloch's theory[59] a stationary state of an electron in a crystal is specified by three components of a vector $\mathbf{k}$, whose rate of change is proportional to the external force acting on the electron. Since the difference in $\mathbf{k}$ between two successive stationary states is in general exceedingly small, the vector $\mathbf{k}$ may be regarded as varying continuously from state to state. The energy E of a state is not a continuous function in $\mathbf{k}$ space, but is discontinuous across certain planes. The positions of these planes are determined by the crystal symmetry, and they divide $\mathbf{k}$ space into *Brillouin zones*, characteristic of the metal. The diamagnetism of the metal, including its anisotropy, depends on the relation of the valency electrons to the nearest zone boundary. The electrons fill up a region in $\mathbf{k}$ space whose outer boundary (sharp only at absolute zero) is a surface of constant energy, called the *surface of Fermi distribution*. If k is the distance in $\mathbf{k}$ space measured normal to a plane across which the energy is discontinuous, the diamagnetic susceptibility is dependent on the value of $\partial^2 E/\partial k^2$, and will be large when $\partial^2 E/\partial k^2$ is large. The Brillouin zones can be located experimentally by the fact that their boundaries correspond to planes giving strong X-ray reflections[25]. It so happens that in the γ-alloys the first important Brillouin zone is a polyhedron having 36 sides at a distance of $3\sqrt{2\pi/a}$ from the origin and is very nearly a sphere. It can contain exactly 90 electrons per unit cell, while the inscribed sphere can contain 80. Now in these alloys the lowest

group of energy levels is almost completely filled by 84 valency electrons per unit cell; that is to say, the surface of Fermi distribution, which is approximately spherical, must lie at all points very close to the Brillouin zone and $\partial^2 E/\partial k^2$ in the immediate neighbourhood of certain crystal planes has a value many times larger than its value for free electrons—37 times as large for {411}, 24 times for {330}.

The bismuth structure is based on a rhombohedral lattice

$$(\alpha = 57° 16', \quad a = 4{\cdot}74 \text{ A.})$$

containing two atoms, each of which has five valency electrons. There is an important Brillouin zone bounded by planes corresponding to the $\{1\bar{1}0\}$ $\{22\bar{1}\}$ and {221}. Such a zone will be almost exactly filled by the ten electrons in the unit cell of bismuth. Jones[25] discusses the shape of the surface of Fermi distribution, and shows that $\chi_{\|}$ (parallel to the principal axis) is mainly due to those parts of the Fermi surface which overlap into the second main zone in the region of the $\{1\bar{1}0\}$ planes, while $\chi_{\perp}$ arises from the parts of the Fermi surface lying within the first zone. The overlap is exceedingly small, approximately 0·0013 electron per atom, but it gives a value of $\chi_{\|}$ of the right order. It follows that if the number of electrons per atom could be decreased, the quantity $\chi_{\perp}/\chi_{\|}$ would increase, but if the number of electrons per atom were increased, $\chi_{\perp}/\chi_{\|}$ would decrease. This variation in the Fermi distribution surface can be achieved, without destroying the structure, by dissolving in the Bi small quantities of Sn, Pb, Te or Se. This was done by Goetz and Focke[13] whose results agree well with the theoretical expectations. The effect increased with lowering of temperature, with the result that for high concentration of an electronegative impurity at sufficiently low temperatures $\chi_{\perp}/\chi_{\|}$ became infinite (i.e. $\chi_{\|} = 0$) and then negative ($\chi_{\|} > 0$). The crystal was then diamagnetic in one direction and paramagnetic in another[16].

A similar co-existence of dia- and para-magnetism has been observed in single crystals of antimony-tin alloys at room temperatures[20], although in this case the value of $\chi_{\perp}/\chi_{\|}$ for pure Sb is less than unity ($\chi_{\perp} = -0{\cdot}55 \times 10^{-6}$, $\chi_{\|} = -1{\cdot}42 \times 10^{-6}$ at 290° K.[56]).

The theory is well supported by the fact that, on melting, the high susceptibilities of the γ-alloys and of Bi, Sb, etc., disappear[41]. The actual anisotropy of Bi and of Sb at room temperatures would seem to be, in a sense, accidental, since $\chi_{\|}$ and $\chi_{\perp}$ vary independently with temperature and with alloying[55]. The theory has not, as yet, been able to give an adequate explanation of temperature dependence[55] or of variation of susceptibility with particle size. The latter has been instanced by Goetz[14] as supporting evidence for the existence of some kind of superstructure in the metallic state.

The de Haas-van Alphen effect[19, 57]. At room temperatures the susceptibility of Bi, like that of all other uniaxial diamagnetic compounds, is independent of field strength and is also constant for all directions in the basal plane. At low temperatures (below 40° K. for fields up to 10,000 gauss) there is a complicated periodic dependence on field and a further variation with a period of 60° in the basal plane. The effect is very pronounced at 4° K. and persists when the Bi is

alloyed with very small concentrations of Pb, Sn, Te and Sb, although both the position and magnitude of the singularities are changed. No such effect could be found for Sb[50]. A partial qualitative theory of this phenomenon has been attempted by Peierls[47], but the dependence on orientation is as yet unexplained.

§ 5. ANISOTROPY OF PARAMAGNETIC CRYSTALS

From table 1 it may be seen that in an isomorphous series of salts the value of the mean susceptibility $\bar{\chi}$ and also the magnitude of the mean anisotropy $\Delta\chi$ are dependent on the nature of the paramagnetic ion[29, 28]. If there is no paramagnetic ion, that is, no ion containing an incomplete electron group or an unpaired spin, the salt is diamagnetic and the small anisotropy existing may be explained as due to differences in the projected electron density distribution in the directions of the principal axes. Such an underlying diamagnetism and diamagnetic anisotropy of course exists in all anisotropic paramagnetic crystals. The effect of crystal structure upon the paramagnetic anisotropy may be seen in the fact that in these isomorphous crystals $\chi_1-\chi_2$ is always greater than $\chi_1-\chi_3$, although the ratio $\chi_1-\chi_2/\chi_1-\chi_3$ is far from constant. It is clear that in order to understand the anisotropy of paramagnetic salts we must know the effect on the carrier of the magnetic moment (1) of the crystalline field, (2) of the magnetic field, in relation both to the intrinsic spin and to the orbital moment.

Quantum theory of paramagnetism. The rigorous quantum mechanical treatment of paramagnetism in ionic crystals, although it has successfully explained many of the experimental results, is too lengthy to be given here. [For references see[66, 54, 3].] It relates the behaviour of the magnetic moment to the spectroscopic state of the ions, and shows the anisotropy of a paramagnetic ion to be dependent on the multiplicity of its lowest level.

In the case of *free* ions, as for example in a paramagnetic gas, having resultant angular momentum j, the $(2j+1)$ m-states will, in the absence of any field, have the same energy. This *degeneracy*, as it is called, may be removed by a magnetic field, which decomposes the single energy level into $2j+1$ nearly equidistant levels. The subsequent distribution of the ions among the levels gives a resultant magnetic moment along the field, which accounts for their paramagnetism. If the ions are not initially free, but are subject to an electric field, the degeneracy in respect of the l (orbital) moment may have been partially or wholly removed by Stark splitting of the levels; in other words, the l moment would have tended to assume a definite orientation in the field which would inhibit its response to an applied magnetic field. The larger the Stark splitting produced by the electric field, the lower the mean susceptibility; if the degeneracy could be completely removed (say by an asymmetric crystalline field) so that the originally degenerate states were completely separated ($h\nu_{\text{sep.}} \gg kT$), then the application of a magnetic field would not lead to any further redistribution and the susceptibility would be zero. Such an electric field is present in the crystalline state, but the degeneracy is not usually completely removed, even in the case of the l moments. The s (spin) moments are left comparatively free to respond to a magnetic field, although

through their coupling with the l moments they may be indirectly affected by the electric field.

The nature of the splitting will depend on the initial states involved, the nature, distances and symmetry of the surrounding atoms. The symmetry of the crystalline field surrounding a paramagnetic ion is not necessarily similar to that of the crystal as a whole, although the fields in isomorphous crystals may be presumed to be similar in character. Unless the crystalline field possesses cubic symmetry, one may expect not only magnetic anisotropy, but also different values of the Curie temperature θ in the expression $\chi(T+\theta)=C$ along the principal axes.

Two groups of salts have been most extensively studied, those of the rare earths[6, 17, 33] and of the first transition series[29, 28].

Rare-earth salts. It is found that in most cases the rare-earth ions behave as if they were free; that is, their magnetic moments may be calculated from Hund's formula (iv) modified to allow for the fact that in some cases (e.g. for Sm and Eu) the multiplet interval is not large compared with kT. This indicates that their l and s moments are strongly coupled and subject to little disturbing interaction on the part of any crystalline field. The behaviour of rare-earth ions in this respect may be traced to the fact that the paramagnetic group of electrons is the incomplete $4f$ sub-shell which is shielded by an outer inert gas configuration of eight electrons of higher quantum number. There is, therefore, no anisotropy due to an unquenched remnant of the l moment. If, however, a rhombic crystal field of definite orientation acts on the rare-earth ion, producing a certain splitting of its lower levels, the Zeeman effect of the spectrum and consequently the magnetic moment of the levels, and the susceptibility, will vary according to the direction of the magnetic field. As Penney and Kynch[48] have pointed out, the ratio $\Delta\chi/\bar{\chi}$ must, at room temperatures, be of the same order as "splitting caused by non-cubic terms in a triply degenerate cubic level"/"overall splitting due to cubic field" (compare figure 2). Actually Krishnan and Mookherji[33] have found values of $\Delta\chi/\bar{\chi}$ for the rare-earth salts $M_2(SO_4)_3.8H_2O$, where M=Sm, Pr, Er, Nd, varying from 21 to 12 per cent. When account is taken of the fact that this anisotropy is a *minimum* value for the group surrounding each rare-earth ion (since there will be several such groups in the unit cell, not in general orientated parallel to each other), it is clear that the departure from cubic symmetry of the field must be quite large. A detailed X-ray study of this isomorphous group of monoclinic crystals would be very desirable.

Salts of the first transition series of elements. In the ions of the transitional groups the incomplete sub-shell of electrons lies outermost and the orbital moments are relatively loosely coupled to the electron spin, and therefore more easily affected by a disturbing electric field[61]. The unquenched remainder of the orbital moment will be anisotropic and therefore indirectly produces the equivalent of a local field which renders the spin contribution to the susceptibility anisotropic also. The anisotropies of salts whose metallic ions have a resultant orbital moment may vary considerably. That of the mineral *siderose* (mainly $FeCO_3$; see table 1) is so large that it has been successfully utilized for the measurement of magnetic fields, in the

direct reading instrument called a gaussmeter[9]. The anisotropy of nickel salts is, however, surprisingly small compared with that of isomorphous iron and cobalt salts. Table 4 gives the values of $\Delta\chi/\bar{\chi}$ (mean anisotropy/mean susceptibility) for the Fe^{++} ($^5D_4$), Co^{++} ($^4F_{9/2}$) and Ni^{++} ($^3F_4$) ions in various salts[20].

Table 4

$\Delta\chi/\chi$ for	$M=Fe^{++}$	Co^{++}	Ni^{++}
$M.SO_4(NH_4)_2SO_4.6H_2O$	12·2%	23·8%	2·7%
$M.SO_4K_2SO_4.6H_2O$	10·7	21·8	3·9
$M.SO_4.7H_2O$	11·8	16·5	2·7

Detailed computations of the effects of crystalline fields on susceptibilities have been made by Van Vleck, Penney and Schlapp[49]. The potential energy of an electron or ion due to the field of the surrounding ions may be expressed as a power series of the coordinates giving distances from the centre of the atom. It is found that many of the experimental results can be explained on the assumption of a predominently cubic field together with a feeble rhombic component, not necessarily coaxial with it. The terms giving rise to the decomposition of the energy levels are

$$\phi=\sum_i\{Ax_i^2+By_i^2-(A+B)z_i^2+D(x_i^4+y_i^4+z_i^4)\} \qquad \text{.....(xiv).}$$

Assuming that the cubic or fourth-order part of (xiv) is dominant, then Bethe's group theory of levels in crystalline fields shows that the Stark splitting of F terms (for which the resultant orbital moment = 3) is arranged either as in figure 2 or as in this figure turned upside down.

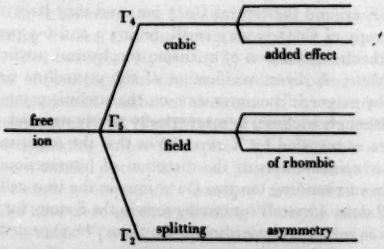


Figure 2. Upright for Ni^{++}; inverted for Co^{++}.

The triple splitting of Γ_4 and Γ_5 is due to the rhombic terms; in a strictly cubic field they would be triply degenerate. If the spin and its coupling to the orbit be included, further decomposition of all the levels occurs, the spin-orbit interaction effect being comparable with that of the rhombic dissymmetry, but entirely subordinate to the separation effected by the cubic field. The general theory of susceptibilities shows that if kT becomes comparable with the separation of the lowest

group of energy levels, Curie's law will cease to be obeyed at low temperatures. Ni^{++} salts conform closely to Curie's law from 300° K. down to 14° K., thus requiring that a very narrow group of levels must lie considerably below all others. This implies that for Ni^{++} the singlet level lies below the two triplet levels (figure 2 upright) and on this assumption the small anisotropy and the conformity to Curie's law may both be understood. Van Vleck[49] has shown theoretically that for the same signs of A, B, D in equation (xiv), as for example in an isomorphous crystalline arrangement, the Stark pattern for Co^{++} will be inverted relative to that for Ni^{++}. The triplet level will therefore be lowest in the rhombic pattern for Co^{++} (figure 2 upside down), giving a multiplicity which is not large compared with kT. This leads to large anisotropy and considerable deviations from Curie's law, behaviour which is typical of Co^{++} salts. Actually the spin is the main cause of the paramagnetism in both Co^{++} and Ni^{++}, though not of its anisotropy. The theory requires that departures of the magnetic moment p_B from the *spin-only* value, equation (iii), should be larger for Co^{++} than for Ni^{++}; this also is confirmed by experiment.

In the hydrated sulphates and selenates, table 1, there is an octahedral arrangement of six negative charges (of the water molecules) around the paramagnetic ion, thus giving a nearly cubic field, with a positive value of D, equation (xiv)[18]. It can be shown that the field of four negative charges at the corners of a tetrahedron about the origin will give a negative value of D, and thus *invert the Stark pattern*. A tetrahedrally surrounded Co^{++} ion may therefore be expected to behave like an octahedrally surrounded Ni^{++} ion, that is, it will have small anisotropy. Krishnan and Mookherji[34] have recently measured the anisotropy of single crystals of $Cs_2[CoCl_4]$ and of $Cs_2[CoCl_4]Cl$, in each of which there is a tetrahedral arrangement of four Cl^- around the central Co^{++} ion, and they have shown that in these salts the anisotropy of Co^{++} is very small, only 3·3 and 6·5 per cent respectively. This is a remarkable verification of quantum-mechanical predictions.

Copper sulphate. A direct verification of the crystalline field theory has also been provided by magnetic measurements on the triclinic crystal $CuSO_4 . 5H_2O$[32]. This crystal, although triclinic, is magnetically nearly uniaxial, and a study of the crystal structure as revealed by X-rays shows that the experimentally determined magnetic axis is coincident with the direction of intersection of two squares of water molecules surrounding the two Cu^{++} ions in the unit cell.

Ions in the S state. Crystals containing ions in the S state, for which the resultant orbital moment is zero, may according to the theory be expected to be magnetically isotropic, even though belonging to non-cubic systems. Salts of the rare-earth ion Gd^{+++} ($^8S_{7/2}$ state) are found to be magnetically isotropic and to obey the Curie law down to liquid helium temperatures within the limits of experimental error. The same is true of Fe^{+++} and Mn^{++} ions ($^6S_{5/2}$)[24]. Krishnan and Banerjee have recently made a detailed study of single crystals of manganous and ferric salts[28]. They have shown that a small anisotropy still exists which is due to three main causes: (1) anisotropy of the diamagnetic constituents of the salt; (2) anisotropy due to the mutual influence of the magnetic spin moments and their non-cubic

arrangement in the crystal lattice; (3) anisotropy of the feeble higher-order Stark splitting of the ^8S levels under the influence of the crystalline electric fields.

The first could be estimated from a study of isomorphous *diamagnetic* salts of Mg and Zn. The effects of the interaction between the spin moments in the crystal were estimated from measurements of the progressive variation of the anisotropy of mixed crystals of

$$Mn(NH_4)_2(SO_4)_2.6H_2O + Mg(NH_4)_2(SO_4)_2.6H_2O$$

and of

$$Mn(NH_4)_2(SO_4)_2{\cdot}6H_2O + Zn(NH_4)_2(SO_4)_2{\cdot}6H_2O$$

with the concentration of the Mn ions. The anisotropy due to this factor was considerably less than that due to the third, the direct effect of the crystalline electric field. Krishnan and Banerjee were able to calculate the small deviations away from the $\chi T = C$ law to be expected along each of the three principal axes, and found the *characteristic temperatures* of the manganous Tutton salts to be of the order of 0·10° K., corresponding to a Stark splitting of the order of $\Delta\nu = 0{\cdot}07$ cm.$^{-1}$ These values, which are deduced from anisotropy measurements made at room temperatures, can be used to predict the entropy-temperature and specific heat temperature curves of these crystals in the neighbourhood of 0·1° K.

In the crystals studied by Krishnan and Banerjee there was considerable *magnetic dilution*, that is, the concentration of the paramagnetic ions in their diamagnetic medium was relatively small. In magnetically concentrated salts such as the anhydrous sulphates and chlorides the spin interaction effects are of more importance and may be expected to lead not only to a somewhat more marked anisotropy, but also to a deviation of the mean susceptibility from the spin-only value.

Complex salts. The behaviour of the metals of the first transition series in complex salts is complicated by the fact that the bonds in complex groups may be covalent rather than ionic. For example, $K_4[Fe(CN)_6]$ is diamagnetic, and so are many other complex salts of paramagnetic ions, the condition being that the *complex ion* involved shall contain an even number of electrons. If the complex ion contains an odd number of electrons, as in $K_3[Fe(CN)_6]$, the mean paramagnetic susceptibility should be that corresponding to one free spin. A considerable amount of experimental work has been carried out on the mean susceptibilities of both diamagnetic and paramagnetic complex compounds. Quantum mechanical explanations of the experimental results have been given by Pauling[46] on the basis of directed wave functions and by Van Vleck[65], who gave alternative theories based on the crystalline field model (although this only applies strictly to ideally ionic crystals) and on the method of molecular orbitals, and showed that all three methods formally predict similar results.

There have been but few attempts to measure the anisotropies of these salts and the variations of individual susceptibilities with temperature. That interesting developments may be expected along these lines is evident from the work of Jackson[24] and Howard[22] on the paramagnetic salt $K_3[Fe(CN)_6]$. Jackson's measurements, covering a temperature range of 290° to 75° K., not only confirmed

the low value of $\bar{\chi}$ reported by previous investigators, but revealed an anisotropy amounting to nearly 100 per cent at low temperatures. By means of calculations using the method of crystalline potentials, but based on general group theory, Howard succeeded in obtaining good quantitative agreement between theory and experiment for the magnitudes, anisotropy and temperature dependence of the principal magnetic susceptibilities of this compound.

REFERENCES

(1) Bates. *Reports on Progress in Physics*, **3**, 181 (1936).
(2) Bhagavantam. *Indian J. Phys.* **4**, 1 (1929).
(3) Bhatnagar and Mathur. *Physical Principles and Applications of Magnetochemistry*, pp. 59–66 (1936).
(4) Bose. *Phil. Mag.* **21**, 1119 (1936).
(5) Bragg. *Nature, Lond.*, **119** (Supplement), 72 (1927).
(6) Cabrera. *Rap. Cons. Phys. Solvay*, **6**, 81 (1930); *C.R. Acad. Sci., Paris*, **205**, 400 (1937).
(7) Cabrera and Fahlenbrach. *Z. Phys.* **89**, 682 (1934).
(8) Clark. *The Electron Structure and Properties of Matter*, p. 221 (1934).
(9) Cotton and Dupouy. *Congr. int. elect.* **3**, 207 (Paris, 1932).
(10) Ehrenfest. *Physica, Eindhoven*, **5**, 388 (1925); *Z. Phys.* **58**, 719 (1929).
(11) Endo. *Sci. Rep. Imp. Univ. Tokyo*, **14**, 479 (1925); **16**, 210 (1927).
(12) *Faraday's Diary*. Edited by L. C. Martin; **5**, 55–148; **6**, 361–93, 409–33, 438–94 (1930).
(13) Ganguli. *Phil. Mag.* **21**, 355 (1936).
Miwo. *Sci. Rep. Tôhoku Univ.* **23**, 242 (1934).
(14) Goetz. *Int. Conf. on Physics, London*, 1934, **2**, 62 (1935).
(15) Goetz and Focke. *Phys. Rev.* **45**, 170 (1934).
(16) Goetz and Focke. *Phys. Rev.* **38**, 1569 (1931).
(17) Gorter. *Arch. Mus. Teyler*, **7**, 183 (1932).
(18) Gorter. *Phys. Rev.* **42**, 437 (1932).
(19) de Haas and van Alphen. *Commun. phys. Lab. Univ. Leiden*, No. 212*a* (1930); No. 220*d* (1932).
(20) Hart. *Phys. Rev.* **52**, 130 (1937).
(21) Honda. *Ann. Phys., Lpz.*, **32**, 1027 (1910).
Owen. *Ann. Phys., Lpz.*, **37**, 657 (1912).
Lessheim. *Curr. Sci.* **5**, 119 (1936).
(22) Howard. *J. chem. Phys.* **3**, 813 (1935).
(23) Hund. *Z. Phys.* **33**, 855 (1925).
(24) Jackson. *Proc. roy. Soc.* A, **140**, 695 (1933).
(25) Jones. *Proc. roy. Soc.* A, **144**, 225 (1934); **147**, 396 (1934).
(26) Krishnan and Banerjee. *Curr. Sci.* **3**, 548 (1935).
(27) Krishnan and Banerjee. *Philos. Trans.* A, **234**, 265 (1935).
(28) Krishnan and Banerjee. *Philos. Trans.* A, **235**, 343 (1936).
(29) Krishnan, Chakravorty and Banerjee. *Philos. Trans.* A, **232**, 103 (1934).
(30) Krishnan and Ganguli. *Nature, Lond.*, **139**, 155 (1937).
(31) Krishnan, Guha and Banerjee. *Philos. Trans.* A, **231**, 235 (1933).
(32) Krishnan and Mookherji. *Phys. Rev.* **50**, 860 (1936).
Ganguli. *Z. Kristallogr.* **93**, 42 (1936).
(33) Krishnan and Mookherji. *Nature, Lond.*, **140**, 549 (1937).
(34) Krishnan and Mookherji. *Phys. Rev.* **51**, 528, 774 (1937).
(35) Krishnan and Raman. *Proc. roy. Soc.* A, **115**, 549 (1927).
(36) London. *C.R. Acad. Sci., Paris*, **205**, 28 (1937); *J. chem. Phys.* **5**, 837 (1937); *J. Phys. Radium*, **8**, 397 (1937).
(37) Lonsdale. Unpublished data.

(38) LONSDALE. *Proc. roy. Soc.* A, **159**, 149 (1937).
(39) LONSDALE. *Nature, Lond.*, **137**, 826 (1936).
(40) LONSDALE and KRISHNAN. *Proc. roy. Soc.* A, **156**, 597 (1936).
(41) MOTT and JONES. *Properties of Metals and Alloys*, p. 211 (1936).
(42) NILAKANTAN. *Phys. Rev.* **52**, 383 (1937).
(43) OXLEY. *Philos. Trans.* A, **214**, 109 (1914).
(44) Pascal's work is well summarized by Bhatnagar and Mathur (ref. (3)), pp. 67–93.
(45) PAULING. *J. chem. Phys.* **4**, 673 (1936).
(46) PAULING. *J. Amer. chem. Soc.* **53**, 1367 (1931).
PAULING and HUGGINS. *Z. Kristallogr.* **87**, 205 (1934).
(47) PEIERLS. *Z. Phys.* **80**, 763 (1933); **81**, 186 (1933).
(48) PENNEY and KYNCH. *Nature, Lond.*, **140**, 109 (1937).
(49) PENNEY and SCHLAPP. *Phys. Rev.* **41**, 194 (1932).
VAN VLECK. *Phys. Rev.* **41**, 208 (1932).
SCHLAPP and PENNEY. *Phys. Rev.* **42**, 666 (1932).
PENNEY. *Phys. Rev.* **43**, 485 (1933).
(50) RABI. *Phys. Rev.* **29**, 174 (1927).
(51) RAMAN. *Nature, Lond.*, **123**, 945; **124**, 412 (1929).
(52) RAO. *Proc. Indian Acad. Sci.* **1**, 123 (1934).
(53) ROBERTSON and WOODWARD. *J. chem. Soc.* p. 1817 (1936).
(54) SCHLAPP and PENNEY. *Reports on Progress in Physics*, **2**, 60 (1935).
(55) SHOENBERG and UDDIN. *Proc. roy. Soc.* A, **156**, 687 (1936).
(56) SHOENBERG and UDDIN. *Proc. Camb. Phil. Soc.* **32**, 499 (1936).
(57) SHOENBERG and UDDIN. *Proc. roy. Soc.* A, **156**, 701 (1936).
(58) SOMMERFELD. *Z. Phys.* **21**, 221; **24**, 360 (1923).
(59) SOMMERFELD and BETHE. *Handbuch der Physik*, **24**, part 2 (1934).
(60) STONER. *Magnetism and Matter*, p. 284 (1934).
(61) STONER. *Phil. Mag.* **8**, 250 (1929).
(62) STONER. *Magnetism and Matter*, p. 509 (1934).
(63) SUCKSMITH. *Phil. Mag.* **8**, 158 (1929).
(64) TYNDALL. *Diamagnetism and Magnecrystallic Action*, p. 2 (1888).
(65) VAN VLECK. *J. chem. Phys.* **3**, 807 (1935).
(66) VAN VLECK and PENNEY. *Phil. Mag.* **17**, 961 (1934).
(67) VOIGT and KINOSHITA. *Ann. Phys., Lpz.*, **24**, 492 (1907).
KRISHNAN. *Phys. Rev.* **38**, 833 (1931).
(68) WILSON. *The Theory of Metals*, pp. 64–122 (1936).
(69) WOOSTER. *Z. Krist. Referatenteil*, **3**, 111 (1931). The author gives a critical account of apparent exceptions to this rule.

CAMBRIDGE: PRINTED BY WALTER LEWIS, M.A., AT THE UNIVERSITY PRESS